The Dissertation Committee for Wen Cui
Certifies that this is the approved version of the following dissertation:

**Variable Selection: Empirical Bayes vs. Fully Bayes**

Committee:

Tom Shively, Supervisor

Edward George, Co-supervisor

Betsy Greenberg

William Jefferys

Tom Sager

Lynn Stokes

# Variable Selection: Empirical Bayes vs. Fully Bayes

**by**

**Wen Cui, B.S., M.S.**

**DISSERTATION**

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

**DOCTOR OF PHILOSOPHY**

THE UNIVERSITY OF TEXAS AT AUSTIN

August 2002

To Mom, Dad, Feiqi, Christine, uncle Eric and aunt Ping.

# Acknowledgments

First, I would like to give my deepest gratitude to Dr. Edward I. George, who led me into the Bayesian world, brought this interesting and challenging problem to me and guided me through with great insight, support and encouragement.

I am also thankful to other committee members, Dr. Tom Shively, Dr. Betsy Greenberg, Dr. William Jefferys, Dr. Tom Sager and Dr. Lynne Stokes, for their kind support and valuable suggestions.

I also thank Dr. Alex Chien for his help in using Latex and sharing his insightful thoughts.

My special appreciation goes to my family for their love, understanding and support.

# Variable Selection: Empirical Bayes vs. Fully Bayes

Publication No. _____

Wen Cui, Ph.D.
The University of Texas at Austin, 2002

Supervisors: Tom Shively
Edward George

For the problem of variable selection for the normal linear model, fixed penalty selection criteria such as AIC, $C_p$, BIC and RIC correspond to the posterior modes of a hierarchical Bayes model for various fixed hyperparameter settings. Adaptive selection criteria obtained by empirical Bayes estimation of the hyperparameters have been shown by George and Foster (2000) to improve on these fixed selection criteria. In this research, we study the potential of alternative fully Bayes methods, which instead margin out the hyperparameters with respect to prior distributions. Several structured prior formulations are considered, and a variety of fully Bayes selection and estimation methods are obtained. Extensive comparisons with their empirical Bayes counterparts suggest that the empirical Bayes methods perform extremely well in spite of their know inadmissibility.

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

The variable selection (or subset selection) problem arises when one wants to model the relationship between a dependent variable, $Y$, and a subset of explanatory variables $X_1, X_2 ... X_p$, but is uncertain about which subset to use. When $p$ is small, a common strategy is to compare all $2^p$ possible models and then select the best one according to some criterion such as $R^2$, adjusted $R^2$, AIC (Akaike,1973) and $C_p$ (Mallows, 1973) – see Hocking (1976), Berk (1977), and Miller (1984,1990) for a discussion and comparisons of these criteria. The selection problem is further complicated when $p$ is large so that comparison of all $2^p$ possible models is computationally prohibitive. In this case, a preliminary strategy is first needed to restrict attention to a smaller, manageable subset of models. Popular algorithms for so reducing the size of the model space, but by no means all, are Efroymson (1960) on stepwise regression, Hocking and Leslie (1967), Furnival and Wilson (1974), Edwards and Havranek (1987) and Narendra and Fukunaga (1977) on branch-and-bound type algorithms. For a comprehensive review of the variable selection problem, see George (2000) and the references therein.

Although classical variable selection techniques are still popular and

used in practice, research continues into the development of new algorithms and selection criteria. In particular, Bayesian approaches to variable selection are beginning to flourish. The attraction of the Bayesian approach includes its ability to incorporate prior information about the model structure and parameters, its probabilistic coherence and its unified formulation of the problem.

The essentials of the Bayesian approach are as follows. The variable of interest, $Y$, is assumed to have a distribution with density $f(Y \mid \theta)$, where $\theta$ is an unknown parameter that fully determines the distribution. It is further assumed that $\theta$ is a random variable with a prior probability distribution $\pi(\theta \mid \lambda)$, where $\lambda$ represents unknown hyperparameters that fully determine the prior. If $\lambda$ were known, the model would then be selected on the basis of the posterior distribution of $\theta$, namely $\pi(\theta \mid Y, \lambda)$. However, $\lambda$ is typically unknown in meaningful Bayesian formulations of the variable selection problem. The main focus of this dissertation is an investigation and comparison of two approaches for dealing with the case of unknown $\lambda$. These two approaches will be referred to as the Empirical Bayes (EB) and the Fully Bayes (FB) approaches. The EB approach for variable selection, which was developed by George and Foster (2000), essentially entails estimating $\lambda$ from the data. The FB approach, which we are developing here, essentially entails putting a hyperprior on $\lambda$ and integrating it out.

We consider the canonical problem of variable selection for the normal linear model. That is, suppose we have $n$ observations on a dependent variable

$Y$ and $p$ independent variables $X = (X_1, ..., X_p)$, which satisfy

$$Y = X\beta + \epsilon, \tag{1.1}$$

where $\epsilon \sim N_n(0, \sigma^2 I)$ and $\beta = (\beta_1, ..., \beta_p)'$ are the unknown coefficients, an unknown portion of which are nonzero. The goal of variable selection is to identify the unknown subset of nonzero $\beta$s. Letting $\gamma = 1, 2, ..., 2^p$ index all $2^p$ possible submodels, each submodel can be expressed as

$$Y = X_\gamma \beta_\gamma + \epsilon,$$

where $\epsilon \sim N_n(0, \sigma^2 I))$, $\beta_\gamma$ is a $q_\gamma \times 1$ vector ($q_\gamma$ is the number of variables in the model) and $X_\gamma$ is a $n \times q_\gamma$ matrix whose columns are the $q_\gamma$ variables in the model. We assume that $\beta_\gamma$ comes from a normal distribution, $p(\beta_\gamma \,|\, \gamma, c)$, indexed by $\gamma$ and a hyperparameter $c$, and that $\gamma$ comes from a distribution, $p(\gamma \,|\, \omega)$, indexed by a hyperparameter $\omega$.

George and Foster (2000) proposed two EB criteria, $C_{MML}$ and $C_{CML}$, which select models corresponding to posterior modes. In $C_{MML}$ the hyperparameters $c$ and $\omega$ are estimated by the MLE (Maximum Likelihood Estimate) of the marginal likelihood function $L(c, \omega \,|\, Y)$. In $C_{CML}$ they are estimated by the MLE of the conditional likelihood function, $L(c, \omega, \gamma \,|\, Y)$. George and Foster (2000) showed that the EB criteria ($C_{CML}$ and $C_{MML}$) can deliver better performance than other classical criteria such as $AIC/C_p$, $BIC$, and $RIC$ over a much wider portion of the model space. To evaluate performance they used expected predictive loss, which is defined as $E\{[X\hat{\beta}(\hat{\gamma}) - X\beta]'[X\hat{\beta}(\hat{\gamma}) - X\beta]\}$.

An alternative FB treatment of the problem is to instead put priors on the hyperparameters, $c$ and $\omega$, and then integrate them out to obtain the marginal posterior, $\pi(\gamma \mid Y)$. As with $C_{MML}$ and $C_{CML}$, the mode of this marginal posterior could then be used to select the model. According to the complete class theorems of decision theory (Berger 1985), any admissible estimator will be a Bayes procedure or a limit of Bayes procedures for suitable loss functions. Because EB procedures are not strictly Bayes, in the sense that they are not obtained through priors or limits of priors on the hyperparameters, one might expect that EB procedures can be improved upon by FB procedures. One the main themes of this dissertation is to investigate the extent to which such improvement can be obtained by computationally feasible FB procedures.

Both the EB and FB approaches are essentially methods for obtaining a posterior over the model space. Because these posteriors can be used in various ways to make inference, we investigate and compare three forms of FB and EB procedures. We will refer to these three forms as Bayes least-squares (Bayes selection followed by least-squares estimation of $\beta$), Bayes posterior mean (Bayes selection followed by posterior-mean estimation of $\beta$) and model averaging. As will be seen, some of my findings are as follows: Simulation evidence suggests that EB procedures based on $C_{MML}$ perform extremely well in all three cases. Procedures based on $C_{MML}$ tend to choose larger models than those based on $C_{CML}$ or $C_{FB}$(notation of the FB criterion). The FB posterior distributions of $\gamma$ are typically multimodal resulting in both instability and computational difficulties. This is especially problematic for the FB model

averaging approach where Markov Chain Monte Carlo is needed for posterior computation.

Surprisingly, in terms of expected predictive loss, the FB selection procedures did not perform as well as the $C_{MML}$ procedures. This may in part be explained by the fact that posterior mode estimates are more appropriate for 0-1 loss functions. Unfortunately, it is prohibitively expensive to use 0-1 loss for simulation evaluation, since the probability of picking exactly the correct model is very small, even for good procedures. The advantage of using expected predictive loss is that it nicely summarizes performance in terms of closeness to the correct model and it is easy to understand. But under such squared error loss, posterior mode selection procedures may not even be admissible.

However, we did find that under a uniform prior for $\omega$, the expressions for the $C_{FB}$ and the $C_{CML}$ procedures were very similar for the Bayes least-squares and Bayes posterior-mean approaches. Such similarity may be partially explained by Deely and Lindley (1981) who showed that, when multiplied by (1 - correction term), an EB posterior can be asymptotically equivalent to a corresponding FB posterior. The correction term is very small and is of order $O(n^{-1})$. This reflects the uncertainty of the MLE for the hyperparameters through the second derivative of the log-likelihood.

This dissertation is organized as follows: in chapter 2, various structures of priors for $c$ and $\omega$ are discussed and considered; under the priors, the FB selection criterion is derived and, together with the least-squares estimator

of $\beta$, its performance is compared with the corresponding EB criteria. In chapter 3, the FB posterior mean estimator is derived and compared with EB counterparts. In chapter 4, both EB and FB model averaging estimators of $\beta$ are derived, and intensive comparison and investigation are done via simulations. Finally, in chapter 5, we discuss the questions raised by this research and suggest potential future research directions.

# Chapter 2

# Bayes Least-Squares Procedures

## 2.1 Empirical Bayes selection criteria

To give the background of the EB criteria, we briefly review George and Foster (2000). The problem is to identify the unknown subset of nonzero $\beta$s, which are the unknown coefficients of a normal linear regression model defined in (1.1). Let $\gamma = 1, 2, ..., 2^p$ index the $2^p$ subsets of $X_1, X_2, ..., X_p$. Let $SS_\gamma = \hat{\beta}_\gamma {X_\gamma}' X_\gamma \hat{\beta}_\gamma$, the regression sum of squares of the $\gamma$th model, where $X_\gamma$ is the $n \times q_\gamma$ matrix whose columns are the $q_\gamma$ variables included in the $\gamma$th model. Let $\hat{\beta}_\gamma = ({X_\gamma}' X_\gamma)^{-1} {X_\gamma}' Y$, the least squares estimate of the coefficient $\beta_\gamma$ in the $\gamma$th model. George and Foster (2000) showed that, assuming that $\beta_\gamma$ comes from a normal distribution and that each variable is independently included in the model with the same probability, the ordering of models by posterior probability $p(\gamma \mid Y)$ is the same as the ordering of models by a penalized sum of squares criterion of the form

$$SS_\gamma / \hat{\sigma}^2 - F\, q_\gamma \tag{2.1}$$

The priors on $\beta_\gamma$ and $\gamma$ under the assumption above can be expressed in the form

$$p(\beta_\gamma, \gamma \mid c, \omega) = p(\beta_\gamma \mid \gamma,\ c) p(\gamma \mid \omega) \tag{2.2}$$

where

$$p(\beta_\gamma \mid \gamma,\ c) = N_{q_\gamma}(0, c\sigma^2((X_\gamma{}'X_\gamma)^{-1})),\ c > 0, \tag{2.3}$$

and

$$p(\gamma \mid \omega) = \omega^{q_\gamma}(1-\omega)^{p-q_\gamma},\ \omega \in (0,1). \tag{2.4}$$

Under these priors, the posterior of $\gamma$ is

$$p(\gamma \mid Y, c, w) \propto \exp\left\{\frac{c}{2(1+c)}\left[SS_\gamma/\sigma^2 - F(c,w)\, q_\gamma\right]\right\}, \tag{2.5}$$

where

$$F(c,w) = \frac{1+c}{c}\left\{2\log\frac{1-w}{w} + \log(1+c)\right\}. \tag{2.6}$$

The ordering result of George and Foster (2000) follows from the fact that given $Y$, $p(\gamma \mid Y, c, w)$ is increasing in

$$SS_\gamma/\sigma^2 - F(c,w)\, q_\gamma. \tag{2.7}$$

It can be shown that the selection criteria AIC (Akaike,1973), $C_p$ (Mallows, 1973), BIC (Schwarz, 1978) and RIC (Foster and George, 1994) can all be expressed as penalized regression sum of squares criteria of the form (2.7), in which $F$ is the corresponding fixed dimensionality penalty. They all select the models that maximize (2.7). If one chooses $c$ and $\omega$ such that $F(c,\omega) = 2$, $\log n$ or $2\log p$, the highest posterior models then correspond to the best models selected by $AIC/C_p$, $BIC$, or $RIC$.

As alternatives to these fixed penalty criteria, George and Foster (2000) proposed two EB criteria, $C_{MML}$ and $C_{CML}$, in which the dimensionality penalties depend on the data and effectively adapt to the model information contained in the data. They further showed that these adaptive penalty EB

criteria can achieve better performance under predictive loss than the fixed penalty criteria. The $C_{MML}$ criterion selects the model that maximizes

$$SS_\gamma/\sigma^2 - F(\hat{c}, \hat{w})\, q_\gamma \tag{2.8}$$

where $\hat{c}$ and $\hat{\omega}$ are the estimators of $c$ and $\omega$ that maximize the marginal likelihood

$$\begin{aligned} L(c, w \mid Y) &\propto \sum_\gamma p(\gamma \mid w)\, p(Y \mid \gamma, c) \\ &\propto \sum_\gamma w^{q_\gamma}(1-w)^{p-q_\gamma}(1+c)^{-q_\gamma/2} \exp\left\{\frac{c\, SS_\gamma}{2\sigma^2(1+c)}\right\} \end{aligned} \tag{2.9}$$

Unfortunately, this maximization can be computationally expensive due to marginalizing over $\gamma$. Although in the special case $X = I$, the computation can be simplified and quite straightforward, it is not feasible to compute $C_{MML}$ for large $p$ when $X$ is nonorthogonal. As a computable approximation of $C_{MML}$, the $C_{CML}$ criterion instead selects the model that maximizes (2.8) in which $\hat{c}$ and $\hat{\omega}$ are the estimators of $c$ and $\omega$ that maximize the conditional likelihood

$$\begin{aligned} L^*(c, w, \gamma \mid Y) &\propto p(\gamma \mid w) p(Y \mid \gamma, c) \\ &\propto w^{q_\gamma}(1-w)^{p-q_\gamma}(1+c)^{-q_\gamma/2} \exp\left\{\frac{c\, SS_\gamma}{2\sigma^2(1+c)}\right\} \end{aligned} \tag{2.10}$$

George and Foster (2000) showed that there is a trade-off in the performance and computation simplicity between $C_{MML}$ and $C_{CML}$. $C_{MML}$ offers better performance than $C_{CML}$ in terms of predictive loss but does not have

a closed form; $C_{CML}$ is more attractive in terms of easy computation and an informative closed form. Their simulation results showed that both $C_{MML}$ and $C_{CML}$ delivered better performance over a much wider portion of the model space than the other fixed penalty criteria mentioned above.

For our comparisons with the FB least-squares procedures below, we consider the EB least-squares procedures which, after a model is selected by $C_{CML}$ or $C_{MML}$, estimates the coefficients of selected model selected by the least-squares, namely $\hat{\beta}_\gamma = (X_\gamma{}' X_\gamma)^{-1} X_\gamma{}' Y$.

## 2.2 Fully Bayes variable selection

The formulation of a FB least-squares procedure is straightforward: hyperpriors are chosen for $c$ and $\omega$ and then these two hyperparameters are integrated out. The model corresponding to the highest poster probability is selected, and then the coefficients for the selected model are estimated by the least-squares estimates. We begin with a discussion of the choice of the hyperpriors.

### 2.2.1 Priors on $\omega$ and c

What priors to use should ideally depend on how much we know about the hyperparameters or the structure of the underlining true model. For example, if we believe that a parsimonious model with large coefficients is more probable, we might choose a prior on $\omega$ that puts more weight on small $\omega$, and a prior on $c$ that puts more weight on large $c$. When no meaningful information

is available, automatic "noninformative" priors will be natural choices.

A simple and popular noninformative prior for $\omega$ is $Uniform(0,1)$. The uniform prior was recommended by Bayes and Laplace based on the principle of insufficient reason, "when there is no evidence to the contrary, all possibilities should be given equal priori weight" (see Novick and Hall (1965), p1107). Geisser (1984) argued that when it's presumed that there is no prior information, the uniform prior is more compelling than the others, such as $\pi(\omega) \propto \omega^{-\frac{1}{2}}(1-\omega)^{-\frac{1}{2}}$, the $Beta(\frac{1}{2}, \frac{1}{2})$ distribution. For more discussions about $\pi(\omega) = \omega^{-\frac{1}{2}}(1-\omega)^{-\frac{1}{2}}$ and the corresponding rules or procedures for obtaining such a prior, one can refer to Jeffreys (1961), Box and Tiao (1973), Akaike (1978), Bernardo (1979), Geisser (1979) and Geisser (1984). A more general prior is $Beta(wa, wb)$. Different combinations of $wa$ and $wb$ will yield substantially different distributions of $\omega$. Therefore, the performance might be very sensitive to the choice of $wa$ and $wb$. We will come back to this again in section 2.3.

For $c$, a natural noninformative prior is the Jeffreys prior. The Jeffreys prior is the square root of the expected Fisher information

$$I(\theta) = -E_\theta \left[ \frac{\partial^2 \log f_\theta(y|\theta)}{\partial \theta^2} \right],$$

where $\theta = (c, \omega)$, and $f_\theta(y \mid \theta)$ is the marginal likelihood of $(c, \omega)$ obtained by marginalizing out the parameters $\beta$ and $\gamma$.

Unfortunately, the marginalizing out of $\gamma$ is not a tractable calculation. To avoid this, we instead consider a conditional Jeffreys prior, a prior of the

form $\pi(c \mid \gamma)$ that depends on $\gamma$. The likelihood of $c$ given $\gamma$ and data $Y$, is $f_c(y|c,\gamma) = \int_{\beta_\gamma} f(y|\beta_\gamma,\gamma)p_{\beta_\gamma}(\beta_\gamma|\gamma,c)d\beta$. Let $\hat{\beta}_\gamma$ be the least-squares estimate of $\beta_\gamma$. Since $\hat{\beta}_\gamma$ is sufficient, $f(y \mid \beta_\gamma,\gamma) = g(y \mid \hat{\beta}_\gamma) \cdot p_{\hat{\beta}_\gamma}(\hat{\beta}_\gamma \mid \beta_\gamma)$. Here, $g(y \mid \hat{\beta}_\gamma)$ is a function that does not depend on $\beta_\gamma$ and, $p_{\hat{\beta}_\gamma}(\hat{\beta}_\gamma \mid \beta_\gamma)$ is the density of the sufficient statistics $\hat{\beta}_\gamma$ given $\beta_\gamma$. Therefore, the conditional Jeffreys prior can actually be obtained from $p(\hat{\beta}_\gamma \mid c,\gamma) = \int_{\beta_\gamma} p_{\hat{\beta}_\gamma}(\hat{\beta}_\gamma \mid \beta_\gamma)p_{\beta_\gamma}(\beta_\gamma \mid \gamma,c)d\beta$.

**Theorem 2.2.1.** *Consider the variable selection problem for the linear model (1.1). Suppose the priors of $\beta_\gamma$ and $\gamma$ are (2.3) and (2.4), respectively. Then the conditional Jeffreys prior of $c$ given $\gamma$ is*

$$\pi(c \mid \gamma) = \frac{\sqrt{\frac{q_\gamma}{2}}}{1+c} \tag{2.11}$$

To prove the theorem, we first recall $f(y \mid \beta_\gamma,\gamma) = g(y \mid \hat{\beta}_\gamma) \cdot p_{\hat{\beta}_\gamma}(\hat{\beta}_\gamma \mid \beta_\gamma)$, where

$$\begin{aligned} g(y \mid \hat{\beta}_\gamma,\gamma) &= |(X_\gamma' X_\gamma)^{-1}|^{\frac{1}{2}}(2\pi)^{-\frac{n-q_\gamma}{2}}(\sigma^2)^{-\frac{n-q_\gamma}{2}} \\ &\quad \cdot \exp\{-\frac{(Y - X_\gamma\hat{\beta}_\gamma)'(Y - X_\gamma\hat{\beta}_\gamma)}{2\sigma^2}\} \end{aligned} \tag{2.12}$$

and

$$p(\hat{\beta}_\gamma \mid \beta_\gamma,\gamma) = N_{q_\gamma}(\beta_\gamma, (X_\gamma' X_\gamma)^{-1}\,\sigma^2). \tag{2.13}$$

Second, the prior of $\beta_\gamma$ is

$$p(\beta_\gamma \mid \gamma, c) = N_{q_\gamma}(0, c\sigma^2\,(X_\gamma' X_\gamma)^{-1}). \tag{2.14}$$

Therefore,

$$p(\hat{\beta}_\gamma \mid c,\gamma) = N_{q_\gamma}(0, (1+c)(X_\gamma' X_\gamma)^{-1}\sigma^2). \tag{2.15}$$

Then the result follows by computing $E\left[\frac{\partial^2 L}{\partial c^2}\right]$, in which

$$L = L(c \mid \hat{\beta}_\gamma, \gamma) = log\left[p(\hat{\beta}_\gamma \mid c, \gamma)\right].$$

A detailed proof can be found in Appendix.

From (2.11), we can see that the prior distribution of $c$ depends on $\gamma$ through the $q_\gamma$. Note that this is an improper prior. Using an improper prior for model selection can be risky in the sense that the posterior probability is not well defined because an arbitrary constant can be associated with the posterior distribution. For the problem of interest in this research, the posterior of the null model does not involve a prior on $c$. Therefore, it is not comparable with the posterior of other models due to the fact that an arbitrary constant will be contained in the posterior probabilities of all the models except the null model. Therefore, a proper prior is desired. Two classes of priors were attempted: priors derived from the conditional Jeffreys priors and conjugate priors. These two classes of priors are natural and simple. Moreover, they lead to very nice closed forms for the posterior.

We first modify the conditional Jeffreys prior to be proper by adjusting the power of $(1+c)$. Since $\sqrt{\frac{q_\gamma}{2}}$ will be canceled out in the normalization, the prior becomes unconditional. Suppose the power of $(1+c)$ is $1+\alpha$ instead of 1, where $\alpha$ is a positive number, then the density function of the prior distribution can be expressed as follows:

$$\pi(c) = \frac{\alpha}{(1+c)^{(1+\alpha)}}. \tag{2.16}$$

Here, $\alpha$ in the prior is another unknown. It can be treated as a hyper-hyperparameter. When $\alpha$ is large, the density function decreases rapidly as $c$ gets large, putting considerable weight on the small values of $c$, especially at zero. Such a prior implies that the coefficients of the variables of the true model are likely to be small. Under such a prior, the FB procedures may favor large models with small coefficients. Therefore, a large $\alpha$ can bias the selection towards models of large sizes and lead to unpleasantly large predictive error when the actual model size is actually small or moderate. On the other hand, as $\alpha$ gets smaller, the relative magnitude of the densities between large $c$ and small $c$ becomes smaller, and the prior will be relatively flat. We would expect that FB procedures would do better in this case than it does in the case of large $\alpha$. In addition, when the prior is small and flat, the likelihood function dominates. Hence, the FB and EB procedures can be very close in selecting the models, unless the likelihood function is multimodal or flat, too. Thus, we expect that a moderately small value of $\alpha$ will be better, and we believe that smaller $\alpha$ is safer than larger $\alpha$, since the FB procedures can be very unstable when $\alpha$ is large. The related simulation results are reported in section 2.4.

Another choice of prior on $c$ is a conjugate prior. Noticing that $c$ actually functions as a scale parameter in a normal distribution, we can choose a conjugate prior for $c$ from the Inverse Gamma family. For example, we can choose an *Inverse Gamma*$(\alpha, b)$ whose density is $\frac{b^\alpha}{\Gamma(\alpha)} \left(\frac{1}{c}\right)^{\alpha+1} e^{-\frac{b}{c}}$. The posterior corresponding to such a prior does not have a closed form due to the difficulty in integrating out $c$. Although a stochastic search can be used

in that case, when $p$ is large, the huge size of the model space will make a stochastic search very inefficient.

From (2.15), we can see that the density of $\hat{\beta}_\gamma$, given $\gamma$ and $c$, is a function of $1+c$. Thus, more tractable expressions may be obtained by putting an Incomplete Inverse Gamma distribution on $(1+c)$ instead of on $c$. It is incomplete in the sense that the domain of $1+c$ is $(1,\infty)$ rather than $(0,\infty)$. Such an Inverse Gamma for $(1+c)$ leads to a very nice form of the posterior. Suppose $(1+c) \sim IIG(\alpha, b)$ on $(1,\infty)$, where IIG stands for Incomplete Inverse Gamma. Since the Jacobian is 1, when we transfer from $1+c$ to $c$, the density function of $c$ is the same as the density of $1+c$.

**Theorem 2.2.2.** *Suppose* $(1+c) \sim IIG(\alpha, b)$ *on* $(1,\infty)$. *Then the probability distribution of* $c$ *has density*

$$\pi(c) = M\frac{1}{(1+c)^{\alpha+1}}e^{-\frac{b}{1+c}}, \tag{2.17}$$

*where* $M = \frac{b^\alpha}{\int_{t=0}^{t=b} t^{\alpha-1}e^{-t}dt}$ *and* $c \in (0,\infty)$.

Proof: Let $u = 1+c \sim IIG(\alpha, b)$ on $(1,\infty)$, then the density of $c$ is

$$\pi(c) = M(1+c)^{-\alpha-1}\exp\{-\frac{b}{1+c}\}|J|,$$

where $c \in (0,\infty)$ and $|J|$ is the Jacobian whose value is 1. $M$ is the norming constant that can be determined as follow.

Let

$$\int_0^\infty \pi(c)dc = \int_0^\infty M(1+c)^{-\alpha-1}\exp\{-\frac{b}{1+c}\}dc = 1.$$

Let $t = \frac{1}{1+c}$, then we have

$$\begin{aligned} \int_0^\infty \pi(c)dc &= \int_0^1 Mt^{\alpha-1}\exp\{-bt\}dt = 1 \\ \Rightarrow \quad M &= \frac{1}{\int_0^1 t^{\alpha-1}\exp\{-bt\}dt} \\ &= \frac{b^\alpha}{\int_{t=0}^{t=b} t^{\alpha-1}e^{-t}dt} \end{aligned} \tag{2.18}$$

The advantages of choosing an Inverse Gamma prior on $1 + c$ are: 1) it is more flexible. It is easier to vary the distribution of $c$ to investigate the performance of FB criteria. 2) the computation of the posterior can be easily carried out. As we will see in section 2.2.2, this prior leads to a very nice closed form for the posterior, which facilitates comparison with EB criteria, especially $C_{CML}$, easily. More pleasantly, (2.16) is actually a special case of (2.17): when $b = 0$, (2.17) reduces to (2.16).

However, the prior introduces two more parameters which must be dealt with. Although these could be estimated by EB or FB methods, the computational difficulty and complexity seem to offset any potential benefits. Instead, we simply choose to assign fixed numbers to them. Our simulations in section 2.4 show that the FB procedures are relatively robust to small changes in the two parameters, especially in $b$. $\alpha$ affects the FB procedure here in the same fashion as discussed in (2.16).

### 2.2.2 Fully Bayes selection criteria

In this section, we derive general FB selection criteria under a *Beta* prior on $\omega$ and an *Incomplete Inverse Gamma* prior on $c$.

**Theorem 2.2.3.** *Consider the variable selection problem for the linear model (1.1). Assume that $X$ is orthogonal and $\sigma$ is known. Let $\gamma$ index subsets of $X_1, X_2, ..., X_p$. $\gamma = 1$ corresponds to the Null set. Let the priors on $\beta_\gamma$, $p(\beta_\gamma \mid \gamma, c)$ be (2.3) and the prior on $\gamma$, $p(\gamma \mid \omega)$ be (2.4). Suppose prior on $\omega$ is $Beta(wa, wb)$ and prior on $c$ is (2.17), then the posterior of $\gamma$ is*

$$\pi(\gamma \mid y) = K \cdot M \cdot e^{\frac{SS_\gamma}{2\sigma^2}} (b + \frac{SS_\gamma}{2\sigma^2})^{-\frac{q_\gamma}{2}-\alpha} \Gamma(\frac{q_\gamma}{2} + \alpha) F_\gamma \pi(\gamma) \tag{2.19}$$

*for $\gamma = 2, 3, ..., 2^p$ and*

$$\pi(\gamma = 1 \mid y) = K \frac{\Gamma(p + wb)}{\Gamma(p + wa + wb)} \frac{\Gamma(wa + wb)}{\Gamma(wb)}. \tag{2.20}$$

*Here, $SS_\gamma = \hat{\beta}'_\gamma X'_\gamma X_\gamma \hat{\beta}_\gamma$ is the regression sum of squares,*

$$\pi(\gamma) = \frac{\Gamma(q_\gamma + wa)\, \Gamma(p - q_\gamma + wb)}{\Gamma(p + wa + wb)} \frac{\Gamma(wa + wb)}{\Gamma(wa)\, \Gamma(wb)},$$

*and $F_\gamma$ is the cumulative density over the interval $(0, b + \frac{SS_\gamma}{2\sigma^2})$ of a gamma distribution, $Gamma(\frac{q_\gamma}{2} + \alpha, 1)$. $K$ and $M$ are the norming constants, where*

$$K = (2\pi)^{-\frac{n}{2}} (\sigma^2)^{-\frac{n}{2}} \exp\{-\frac{Y'Y}{2\sigma^2}\} \,/\, m(y),$$

*$m(y)$ is the marginal density function of $Y$, and $M$ is defined in (2.18).*

Proof: Let $\pi(\gamma \mid y)$ denote the posterior, then

$$\begin{aligned}
\pi(\gamma \mid y) &\propto \int_\omega \int_c \int_{\beta_\gamma} f(y \mid \beta_\gamma, \gamma) p(\beta_\gamma \mid \gamma, c) \pi(c) p(\gamma \mid \omega) \pi(\omega) d\beta_\gamma \, dc \, d\omega \\
&= \int_c \int_{\beta_\gamma} g(y \mid \hat{\beta}_\gamma, \gamma) p(\hat{\beta}_\gamma \mid \beta_\gamma, \gamma) p(\beta_\gamma \mid \gamma, c) \pi(c) \, d\beta_\gamma \, dc \\
&\quad \cdot \int_\omega p(\gamma \mid \omega) \pi(\omega) d\omega \\
&= g(y \mid \hat{\beta}_\gamma, \gamma) \int_c p(\hat{\beta}_\gamma \mid \gamma, c) \pi(c) \, dc \int_\omega p(\gamma \mid \omega) \pi(\omega) d\omega,
\end{aligned}$$

where $\hat{\beta}_\gamma$ is the least-square estimate of $\beta_\gamma$, $g(y \mid \hat{\beta}_\gamma, \gamma)$ is defined in (2.12), $p(\hat{\beta}_\gamma \mid \beta_\gamma, \gamma)$ is defined in (2.13) and $p(\hat{\beta}_\gamma \mid \gamma, c)$ is defined in (2.15). Then,

$$\begin{aligned} \pi(\gamma \mid y) \quad \propto \quad & (2\pi)^{-\frac{n}{2}} (\sigma^2)^{-\frac{n}{2}} \exp\{-\frac{Y'Y}{2\sigma^2}\}\, M\, \exp\{\frac{SS_\gamma}{2\sigma^2}\} \\ & \cdot \int_0^\infty (1+c)^{-\frac{q_\gamma}{2}-\alpha-1} \exp\{-\frac{b+\frac{SS_\gamma}{2\sigma^2}}{1+c}\} dc \cdot \pi(\gamma). \end{aligned} \tag{2.21}$$

Here, $M$ is (2.18) and

$$\pi(\gamma) = \frac{\Gamma(q_\gamma + wa)\,\Gamma(p - q_\gamma + wb)}{\Gamma(p + wa + wb)} \frac{\Gamma(wa+wb)}{\Gamma(wa)\,\Gamma(wb)} \tag{2.22}$$

Let $m(y)$ be the marginal of $Y$ and $K = (2\pi)^{-\frac{n}{2}} (\sigma^2)^{-\frac{n}{2}} \exp\{-\frac{Y'Y}{2\sigma^2}\} \,/\, m(y)$. Then the posterior is

$$\pi(\gamma \mid y) = K \cdot M \cdot e^{\frac{SS_\gamma}{2\sigma^2}} \cdot \int_0^\infty (1+c)^{-\frac{q_\gamma}{2}-\alpha-1} \exp\{-\frac{b+\frac{SS_\gamma}{2\sigma^2}}{1+c}\} dc \cdot \pi(\gamma)$$

Let $s = \frac{1}{1+c}$,

$$\pi(\gamma \mid y) = K \cdot M \cdot e^{\frac{SS_\gamma}{2\sigma^2}} \cdot \int_{s=0}^{s=1} s^{\frac{q_\gamma}{2}+\alpha-1} \exp\{-(b + \frac{SS_\gamma}{2\sigma^2})s\} ds \cdot \pi(\gamma)$$

Let $t = (b + \frac{SS_\gamma}{2\sigma^2})s$,

$$\begin{aligned} \pi(\gamma \mid y) \quad &= \quad K \cdot M \cdot e^{\frac{SS_\gamma}{2\sigma^2}} \frac{\Gamma(\frac{q_\gamma}{2}+\alpha)}{(b+\frac{SS_\gamma}{2\sigma^2})^{\frac{q_\gamma}{2}\alpha}} \int_{t=0}^{t=b+\frac{SS_\gamma}{2\sigma^2}} \frac{t^{\frac{q_\gamma}{2}+\alpha-1} e^{-t}}{\Gamma(\frac{q_\gamma}{2}+\alpha)} dt \cdot \pi(\gamma) \\ &= \quad K \cdot M \cdot e^{\frac{SS_\gamma}{2\sigma^2}} (b + \frac{SS_\gamma}{2\sigma^2})^{-\frac{q_\gamma}{2}-\alpha} \Gamma(\frac{q_\gamma}{2}+\alpha)\, F_\gamma\, \pi(\gamma) \end{aligned}$$

Here, $F_\gamma$is the cumulative density on $(0, b + \frac{SS_\gamma}{2\sigma^2})$ of the gamma distribution, $Gamma(\frac{q_\gamma}{2}+\alpha, 1)$, i.e.,

$$F_\gamma = \int_{t=0}^{t=b+\frac{SS_\gamma}{2\sigma^2}} \frac{t^{\frac{q_\gamma}{2}+\alpha-1} e^{-t}}{\Gamma(\frac{q_\gamma}{2}+\alpha)} dt.$$

If $\gamma = 1$, i.e, the model is the Null model, then $Y \mid \gamma = 1 \sim N(0, \sigma^2 I)$, which does not depend on $\beta$ and $c$, and

$$\begin{aligned}
\pi(\gamma = 1) &= \int_\omega \omega^{q_\gamma}(1-\omega)^{p-q_\gamma} \frac{\Gamma(wa+wb)}{\Gamma(wa)\,\Gamma(wb)} \omega^{wa-1}(1-\omega)^{wb-1} d\omega|_{\gamma=1} \\
&= \frac{\Gamma(wa+1)\,\Gamma(p+wb-1)}{\Gamma(p+wa+wb)} \frac{\Gamma(wa+wb)}{\Gamma(wa)\,\Gamma(wb)} \\
&= \frac{\Gamma(p+wb)}{\Gamma(p+wa+wb)} \frac{\Gamma(wa+wb)}{\Gamma(wb)}, \qquad (2.23)
\end{aligned}$$

Since $SS_{\gamma=1} = 0$,

$$\begin{aligned}
\pi(\gamma = 1 \mid y) &= \frac{f(y \mid \gamma = 1)\pi(\gamma = 1)}{m(y)} \\
&= \frac{(2\pi)^{-\frac{n}{2}}(\sigma^2)^{-\frac{n}{2}} \exp\{-\frac{Y'Y}{2\sigma^2}\}}{m(y)} \frac{\Gamma(p+wb)}{\Gamma(p+wa+wb-1)} \frac{\Gamma(wa+wb)}{\Gamma(wb)} \\
&= K \frac{\Gamma(p+wb)}{\Gamma(p+wa+wb)} \frac{\Gamma(wa+wb)}{\Gamma(wb)}.
\end{aligned}$$

If $b = 0$, the posterior corresponding to the prior (2.16) is

$$\pi(\gamma \mid y) = K \cdot M \cdot e^{\frac{SS_\gamma}{2\sigma^2}} \left(\frac{SS_\gamma}{2\sigma^2}\right)^{-\frac{q_\gamma}{2}-\alpha} \Gamma\Big(\frac{q_\gamma}{2}+\alpha\Big)\, F_\gamma\, \pi(\gamma) \qquad (2.24)$$

and $\pi(\gamma = 1 \mid y)$ is the same as (2.20). Here, $K$ is the same as before, $M = \alpha$ and $F_\gamma$ is the cumulative density on $(0,\ \frac{SS_\gamma}{2\sigma^2})$ of gamma$(\frac{q_\gamma}{2}+\alpha,\ 1)$.

The logarithm of (2.19) is proportional to

$$\log M + \frac{SS_\gamma}{2\sigma^2} - (\frac{q_\gamma}{2}+\alpha)\log(\frac{SS_\gamma}{2\sigma^2}) + \log\Gamma\Big(\frac{q_\gamma}{2}+\alpha\Big) + \log F_\gamma + \log\pi(\gamma) \qquad (2.25)$$

and the logarithm of (2.20) is proportional to

$$\log\Gamma(p+wb) - \log\Gamma(p+wa+wb) + \log\Gamma(wa+wb) - \log\Gamma(wb). \qquad (2.26)$$

The FB selection criterion picks the model that maximizes (2.25) together with (2.26).

## 2.3 Empirical Bayes vs. Fully Bayes

In this section, we will compare the FB criterion with the EB criterion. Before we proceed, we will rewrite (2.25) and discuss its interesting similarity to $C_{CML}$.

Surprisingly, if we rewrite (2.25), we obtain an expression of the FB posterior that is very close to $C_{CML}$ and can be compared with $C_{CML}$ side-by-side. For the purpose of comparison, we will use similar notations ($B^*$ and $R^*$) to those ($B$ and $R$)used by George and Foster (2000) in $C_{CML}$. We multiply the log posterior (2.25) by 2 and define the new expression as $C_{FB}$ as follows:

$$C_{FB} = \frac{SS_\gamma}{\sigma^2} - B^*(SS_\gamma/\sigma^2) - R^*(q_\gamma) \tag{2.27}$$

where

$$\begin{aligned} B^*(SS_\gamma/\sigma^2) &= (q_\gamma + 2\,\alpha)\,\log\frac{SS_\gamma}{2\sigma^2} - 2\log\Gamma\Big(\frac{q_\gamma}{2}+\alpha\Big) - 2\log\ F_\gamma - 2\ \log M \\ &= q_\gamma(\,1 + \log_+\frac{SS_\gamma}{\sigma^2 q_\gamma}\,) + \Delta B^*(\frac{SS_\gamma}{\sigma^2}) + \Delta B^*(q_\gamma) - 2\log M \end{aligned} \tag{2.28}$$

where

$$\begin{aligned} \Delta B^*(\frac{SS_\gamma}{\sigma^2}) &= q_\gamma\left[\log\left(\frac{SS_\gamma}{\sigma^2 q_\gamma}\right) - \log_+\left(\frac{SS_\gamma}{\sigma^2 q_\gamma}\right)\right] + 2\alpha\log\left(\frac{SS_\gamma}{2\sigma^2}\right) - 2\log F_\gamma \\ \Delta B^*(q_\gamma) &= q_\gamma\ \left[\ \log(\frac{q_\gamma}{2}) - \frac{2}{q_\gamma}\ \log\ \Gamma\Big(\frac{q_\gamma}{2}+\alpha\Big) - 1\right] \end{aligned}$$

and

$$\begin{aligned} R^*(q_\gamma) &= -2\log(\pi(\gamma)) \\ &= -2\left[\log\frac{\Gamma(q_\gamma+wa)\Gamma(p-q_\gamma+wb)}{\Gamma(p+wa+wb)}+\log\frac{\Gamma(wa+wb)}{\Gamma(wa)\Gamma(wb)}\right] \end{aligned} \tag{2.29}$$

When $wa = wb = 1$, i.e., the prior on $\omega$ is *Uniform*(0,1),

$$\begin{aligned} R^*(q_\gamma) &= -2\left[\log\frac{\Gamma(q_\gamma+1)\Gamma(p-q_\gamma+1)}{\Gamma(p+2)}\right] \\ &= -2\left[\log(q_\gamma)!+\log(p-q_\gamma)!-\log(p+1)!\right] \\ &= -2\left\{(p-q_\gamma)\log(p-q_\gamma)+q_\gamma\log q_\gamma\right\}+\Delta R^* \end{aligned} \tag{2.30}$$

where

$$\Delta R^* = 2\left[(p-q_\gamma)\log(p-q_\gamma)+q_\gamma\log q_\gamma-\log q_\gamma!-\log(p-q_\gamma)!+\log(p+1)!\right].$$

We will see that when the prior on $\omega$ is *Uniform*(0,1), $C_{CML}$ is a very good approximation to $C_{FB}$. $C_{CML}$ from George & Foster (2000) was expressed as follow:

$$C_{\text{CML}} = SS_\gamma/\sigma^2 - B(SS_\gamma/\sigma^2) - R(q_\gamma) \tag{2.31}$$

where

$$B(SS_\gamma/\sigma^2) = q_\gamma\left\{1+\log_+(SS_\gamma/\sigma^2 q_\gamma)\right\}, \tag{2.32}$$

$\log_+(\cdot)$ is the positive part of $\log(\cdot)$, and

$$R(q_\gamma) = -2\left\{(p-q_\gamma)\log(p-q_\gamma)+q_\gamma\log q_\gamma\right\}. \tag{2.33}$$

Both (2.27) and (2.31) are penalized regression sums of squares. In $C_{CML}$, $B(SS_\gamma/\sigma^2)$ is the penalty resulted from estimating $c$ by the conditional maximum likelihood estimator $\hat{c} = (SS_\gamma/\sigma^2 q_\gamma - 1)_+$, where $(\cdot)_+$ is the

positive-part function, and $R(q_\gamma)$ is the penalty resulted from estimating $\omega$ by the conditional maximum likelihood estimator $\hat{w}_\gamma = q_\gamma/p$. Similarly, in $C_{FB}$, $B^*(SS_\gamma/\sigma^2)$ is the penalty due to marginalizing over $c$ and $R^*(q_\gamma)$ is the penalty due to marginalizing over $\omega$. The difference between $C_{FB}$ and $C_{CML}$ is essentially reflected in $\Delta B^*(SS_\gamma/\sigma^2) + \Delta B^*(q_\gamma) - 2\log M + \Delta R^*$.

In $\Delta B^*(SS_\gamma/\sigma^2)$, $q_\gamma \left[\log\left(\frac{SS_\gamma}{\sigma^2 q_\gamma}\right) - \log_+\left(\frac{SS_\gamma}{\sigma^2 q_\gamma}\right)\right]$ can be ignored, unless $SS_\gamma$ is very small. Since $\frac{SS_\gamma}{2\sigma^2} = O(q_\gamma n)$, if the diagonal elements of $X'X$ is $O(n)$ then $\log\left(\frac{SS_\gamma}{2\sigma^2}\right) = O(\log(q_\gamma\, n))$. $-2\log F_\gamma$ achieves values between $[0, g(q_\gamma))$, where $g(q_\gamma)$ is a positive function of $q_\gamma$ that shouldn't be too large unless $SS_\gamma/\sigma^2$ is very small. Therefore, $\Delta B^*(SS_\gamma/\sigma^2)$ is an increasing function in both $q_\gamma$ and $n$ and positive unless $SS_\gamma/\sigma^2$ is smaller than $q_\gamma$. The magnitude can be amplified when $\alpha$ is large.

$\Delta B^*(q_\gamma)$ is a decreasing negative function. Therefore, $\Delta B^*(q_\gamma)$ penalizes the regression sum of squares in the opposite direction. To be more precise, it actually rewards addition of variables. Hence, due to the fact that $\Delta B^*(SS_\gamma/\sigma^2)$ and $\Delta B^*(q_\gamma)$ penalizes $\frac{SS_\gamma}{\sigma^2}$ in the opposite way, the penalty resulting from marginalizing over $c$ in $C_{FB}$ should not differ much from the penalty resulted from approximating $c$ by the conditional maximum likelihood estimate in $C_{CML}$. However, if $\alpha$ is very large, $\Delta B^*(q_\gamma)$ will be significantly dominated by $\frac{2}{q_\gamma}\ \log\ \Gamma\left(\frac{q_\gamma}{2} + \alpha\right)$ so that $C_{FB}$ will heavily favor large models. The consequence is that the performance of the $C_{FB}$ can be hurt when the model is actually parsimonious.

It can be shown that $B^*(SS_\gamma/\sigma^2)$ penalizes $C_{FB}$ by approximately $1 +$

$\log_+ \left(\frac{SS_\gamma}{\sigma^2 q_\gamma}\right) + \frac{2}{q_\gamma} \log\left(\frac{q_\gamma}{2} + \alpha - 1\right)$ for adding a variable. This differs from the penalty for one additional variable in $C_{CML}$ by $\frac{2}{q_\gamma} \log\left(\frac{q_\gamma}{2} + \alpha - 1\right)$, which is positive and decreasing (except when both $\alpha$ and $q_\gamma$ is small) function in $q_\gamma$. Therefore, by marginalizing over $c$, $C_{FB}$ tends to penalize more heavily than $C_{CML}$ for adding a variable. However, unless $\alpha$ is very large, the magnitude will be very small compared with $\log_+ \left(\frac{SS_\gamma}{\sigma^2 q_\gamma}\right)$. Therefore, marginalizing over $c$ with respect to a noninformative prior does not seem to be very different from estimating $c$ from the conditional maximum likelihood function.

It can also be easily seen that $B^*(SS_\gamma/\sigma^2)$ in $C_{FB}$ depends on $\alpha$. Suppose that $b = 0$. When $\alpha$ is small, which says that the prior tends to be flat and small everywhere, both $\Delta B^*(SS_\gamma/\sigma^2)$ and $\Delta B^*(q_\gamma)$ won't vary too much as $q_\gamma$ changes. When $\alpha$ is large, which says that small $c$ is more likely than large $c$, $\Delta B(q_\gamma)$ will be dominated by $\frac{2}{q_\gamma} \log \Gamma\left(\frac{q_\gamma}{2} + \alpha\right)$ and reward $C_{FB}$ for addition of a variable. That is, a prior that favors small $c$ can lead to the FB criterion that favors large models over small models. Therefore, unless one has reason to believe that $c$ is large, one should not choose a large $\alpha$. It can lead to a disaster when $c$ is actually not small. This observation supports what we discussed in section 2.2.1.

The other parameter, $b$, influences $C_{FB}$ through $F_\gamma$ and $M$. When $b$ is small, $F_\gamma$ is dominated by $SS_\gamma$. $C_{FB}$ will be affected very little by the change in $b$. When $b$ is large, $F_\gamma$ will be close to 1 since $SS_\gamma$ is usually large. Therefore, $C_{FB}$ will not be sensitive to the change in $b$ again. Since $M$ does not involve $q_\gamma$, changing $b$ only affects the magnitude but not the shape of $C_{FB}$. In this

sense, $C_{FB}$ is robust to the change of $b$. Simulations in section 2.4 also support this observation.

We will first discuss the behavior of $R^*(q_\gamma)$ for the case when $wa = wb = 1$. In that case, the $R^*(q_\gamma)$ in $C_{FB}$ doesn't differ much from the $R(q_\gamma)$ in $C_{CML}$. Actually, $\Delta R^*$ can be expressed as $2\,S(q_\gamma) + 2\log{(p+1)!}$, where $S(q_\gamma)$ is an upside-down U-shaped function of $q_\gamma$ whose value varies in a small range, and $\Delta R^*$ is dominated by $2\log{(p+1)!}$. Therefore, $R^*(q_\gamma)$ in $C_{FB}$ and $R(q_\gamma)$ in $C_{CML}$ only differs by approximately a constant. Also, it's easy to show that for each additional variable, $R^*(q_\gamma)$ and $R(q_\gamma)$ penalize $C_{FB}$ and $C_{CML}$ respectively by the same amount, $2\log{\frac{p-q_\gamma+1}{q_\gamma}}$.

In addition, we found that $C_{FB}$ was also bimodal overall: the two largest modes are at the two ends. One is close to the null model and the other is at the Full model and, often times, the later is where $C_{FB}$ is maximized. Such bimodality is totally reflected in the $R^*(q_\gamma)$ penalty, which is resulted from integrating out $\omega$. Due to the overall bimodality, the performance of $C_{FB}$ is usually worse when the size of the model is actually around half of the total variables of interest, say $200 \sim 700$. The presence of the bimodality suggests that integrating out $c$ or $\omega$ with respect to the noninformative priors does not give the Full Bayes procedure the ability to distinguish models with many small coefficients from models with a few large coefficients. A technical modification proposed by George and Foster (2000) has improved the performance of $C_{CML}$. This modification helped to improve the performance of $C_{FB}$, and was adopted here, too.

When $wa \neq 1$ and $wb \neq 1$, the $Beta(wa, wb)$ can have vary different shapes as $wa$ and $wb$ vary. These priors weight $\omega$ unequally, and imply that models of certain sizes are believed to be more probable than others. For example, if $wa > 1$ and $wb \leq 1$, the density function of $Beta(wa, wb)$ is increasing in $\omega$, which implies that large models are more probable than small models. Such preference is then translated into the posterior and is reflected accordingly in $R^*(q_\gamma)$: when $wa > 1$ and $wb < 1$, $R^*(q_\gamma)$ is unbalanced upside-down U-shaped, and it penalizes small models more than it penalizes large models. Varying $wa$ and $wb$ can certainly improve the performance of $C_{FB}$ on some portion of the model space, but not on the overall model space. The performance of $FB$ for five typical $Beta$ densities (uniform, increasing function, decreasing function, symmetric unimodal function and U-shaped function) are investigated and compared in section 2.4. As we will see that none of them enables $FB$ to achieve uniformly better performance.

Since $C_{MML}$ does not have a closed form, it can not be compared with $C_{FB}$ analytically. In section 2.4 we will compare the two via the simulations.

## 2.4 Simulations

This section consists of three parts. In the first part, we describe the data generation procedure. In the second part, we investigate the robustness of $C_{FB}$ to different priors on $c$ and $\omega$. In the third part, we discuss and compare the simulation evaluations on the performance of $C_{FB}$, $C_{CML}$ and $C_{MML}$. We will only focus on the case when $X$ is orthogonal. All the proceed-

ing discussions can be applied to nonorthogonal cases. However, when $X$ is nonorthogonal, the simulations can be done in similar ways but will be more computationally intensive.

When $X$ is orthogonal, we set $X = I$ for the simplicity. Then (1.1) is reduced to $Y = \beta + \epsilon$. The same data generation procedure in George and Foster (2000) is adopted here: for each fixed $\alpha$, $b$, $wa$, $wb$ and $q$, $n = p = 1000$, observations of $Y$ were generated by first generating the first q nonzero components in $\beta$ and setting $\beta_{q+1}, \beta_{q+2}, \ldots, \beta_p$ to be zeros, and then adding the independent normal noise $\epsilon$ (i.e., $\epsilon \sim N_p(0, I_p)$) to $\beta$. The same procedures were repeated for $m$ times and the average predictive error loss

$$L\{\beta, \hat{\beta}_\gamma)\} \equiv \{X\hat{\beta}_\gamma - X\beta\}'\{X\hat{\beta}_\gamma - X\beta\} \tag{2.34}$$

were calculated over all the $m$ replications to investigate the robustness of the FB procedure to the different choices of the hyperpriors, or to compare the FB procedure with the EB procedure. Here, $\hat{\beta}_\gamma$ is the least-squares estimate of $\beta_\gamma$.

Before we proceed, we want to point out that for each fixed $q_\gamma$, the posterior (2.19) is monotonically increasing in $SS_\gamma$. Let's rewrite (2.19) as following:

$$\pi(\gamma) \propto H_1(q_\gamma) \cdot H_2(SS_\gamma)$$

where

$$H_1(q_\gamma) = M\, \Gamma(\frac{q_\gamma}{2} + \alpha)\, \pi(\gamma)$$

and

$$H_2(SS_\gamma) = e^{\frac{SS_\gamma}{2\sigma^2}} \left(b + \frac{SS_\gamma}{2\sigma^2}\right)^{-\frac{q_\gamma}{2}-\alpha} F_t(b + \frac{SS_\gamma}{2\sigma^2}, \frac{q_\gamma}{2} + \alpha).$$

$H_1(q_\gamma)$ only depends on $q_\gamma$ and is constant when $q_\gamma$ is fixed. It is easily seen that $H_2(SS_\gamma)$ is a monotonically increasing function in $SS_\gamma$ for each fixed $q_\gamma$. Such a nice property allows us to simplify the computation when X is orthogonal. When X is orthogonal, $SS_\gamma$ is simply the sum of squares of $q_\gamma$ independent variables. Therefore, we only need compute the posterior for the $p+1$ distinct models to maximize (2.19) and the computation is straightforward. The $p+1$ distinct models are the best models of $p$ distinct sizes plus the null model. If $X$ is nonorthogonal, one can use a Metropolis-Hastings algorithm to search the promising model stochastically. Or, one can first reduce the model space to a subset and then apply the selection criteria to the subsets.

### 2.4.1 Robustness to the choices of hyperpriors

Simulations were run to test the sensitivity of $C_{FB}$ to the choices of $\alpha$, $b$, $wa$ and $wb$. It was found that small $h$ is better than large $h$, $C_{FB}$ is very insensitive to the change in $b$, and $C_{FB}$ is more sensitive to the change in $wa$ and $wb$ than to the change in $\alpha$ and $b$.

Simulation 1: Investigate the robustness of $C_{FB}$ to the choice of $\alpha$:

We fix $n = p = 1000$, $c = 5$, $b = 0$, $wa = wb = 1$. Using above data generation procedure, $m = 500$ replications of data were generated for each $q$

| q | 0 | 10 | 25 | 50 | 100 | 200 | 300 | 400 | 500 | 750 | 1000 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $C_{MML}$ | 3.84 | 38.07 | 79.41 | 143.5 | 258.09 | 457.84 | 626.73 | 769.69 | 878.41 | 999.60 | 998.70 |
| $C_{CML}$ | 0.17 | 34.79 | 81.04 | 150.11 | 272.58 | 494.60 | 684.49 | 850.34 | 988.84 | 1244.14 | 1186.44 |
| $C_{FB}$ alpha=1 | 0.13 | 35.09 | 80.92 | 149.66 | 271.87 | 493.77 | 683.94 | 849.56 | 988.16 | 1243.20 | 1186.09 |
| $C_{FB}$ alpha=0.0000001 | 0 | 40.15 | 84.33 | 150.77 | 272.63 | 494.58 | 684.48 | 850.49 | 988.90 | 1244.41 | 1189.39 |
| $C_{FB}$ alpha=0.001 | 0 | 37.22 | 83.14 | 150.54 | 272.63 | 494.58 | 684.48 | 850.49 | 988.90 | 1244.41 | 1189.39 |
| $C_{FB}$ alpha=10 | 0 | 45.55 | 98.89 | 162.47 | 272.51 | 489.19 | 677.90 | 843.44 | 981.69 | 1234.34 | 1162.77 |
| $C_{FB}$ alpha=100 | 0 | 50.57 | 126.94 | 248.56 | 497.50 | 997.24 | 1464.18 | 1917.05 | 2269.05 | 3024.55 | 3130.99 |
| $C_{FB}$ alpha=1000 | 489.6 | 436.87 | 367.84 | 283.23 | 267.38 | 698.63 | 1193.67 | 1697.24 | 2189.58 | 3454.24 | 4687.66 |

Table 2.1: Robustness of $C_{FB}$ to $\alpha$: average losses of $C_{CML}$, $C_{MML}$ and $C_{FB}$. m=500, n=p=1000, c=5, b=0 and $wa = wb = 1$.

of 0, 10, 25, 50, 100, 200, 300, 400, 500, 750 and 1000. $C_{FB}$ was applied to the same data for each $\alpha$ of 0.0000001, 0.001, 1, 10, 100 and 10000. $C_{CML}$ and $C_{MML}$ were also applied to the same data for comparison. The averages losses over all the replications were calculated. As we can see from Figure (2.1) and Table (2.1), $C_{FB}$ is more stable for small $\alpha$ than for large $\alpha$. When $\alpha$ is very small, e.g. $\alpha = 0.0000001$ or $\alpha = 0.001$, $C_{FB}$ is almost the same as $C_{CML}$. However, when $\alpha$ is large, $C_{FB}$ can be much worse than $C_{CML}$ as we can see well from Table (2.1). Such result is consistent with what we have discussed in section (2.3) regarding the influence of $\alpha$ on the FB procedure. The simulation results also show that in no case is $C_{FB}$ better than $C_{MML}$. Since there isn't much difference when $\alpha$ is small, say $\alpha < 10$, we will fixed $\alpha$ at 1 in the rest simulations in this research for simplicity.

Simulation 2: Investigate the robustness of $C_{FB}$ to the choice of $b$.

The procedure is the same as that in "Simulation 1", except, this time, we fix $\alpha$ at 1 and vary $b$. The other parameters, $m$, $n$, $p$, $c$, $wa$, $wb$ and $q$ are the same. $C_{FB}$ was repeated on the same data for $b$= 0, 1, 10, 100 and 10000. The EB criteria ($C_{CML}$ and $C_{MML}$) were included for comparison. Figure (2.2) and Table (2.2) show that the performance of $C_{FB}$ varies very little when $b$

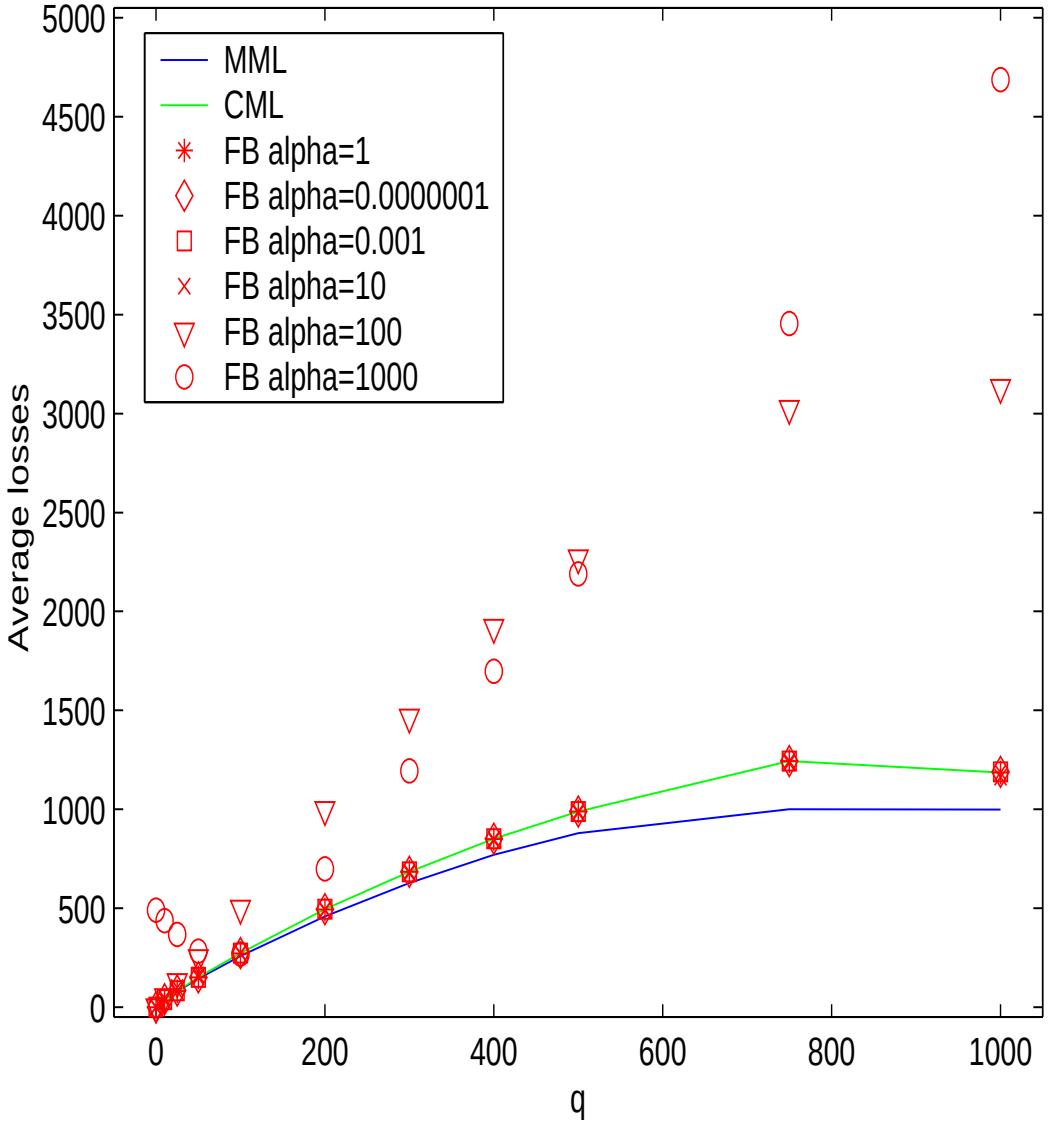


Figure 2.1: Robustness of $C_{FB}$ to $\alpha$: average losses of $C_{CML}$, $C_{MML}$ and $C_{FB}$. m=500, n=p=1000, c=5, b=0 and $wa = wb = 1$.

| q | 0 | 10 | 25 | 50 | 100 | 200 | 300 | 400 | 500 | 750 | 1000 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $C_{MML}$ | 3.84 | 38.07 | 79.41 | 143.50 | 258.09 | 457.84 | 626.73 | 769.69 | 878.41 | 999.60 | 998.70 |
| $C_{CML}$ | 0.17 | 34.79 | 81.04 | 150.11 | 272.58 | 494.60 | 684.49 | 850.34 | 988.84 | 1244.14 | 1186.44 |
| $C_{FB}$ b=0 | 0.13 | 35.09 | 80.92 | 149.66 | 271.87 | 493.77 | 683.94 | 849.56 | 988.16 | 1243.20 | 1186.09 |
| $C_{FB}$ b=1 | 0.13 | 34.93 | 80.80 | 149.63 | 271.87 | 493.77 | 683.94 | 849.56 | 988.16 | 1243.20 | 1186.09 |
| $C_{FB}$ b=10 | 1.14 | 33.76 | 80.15 | 149.57 | 271.87 | 493.77 | 683.94 | 849.56 | 988.16 | 1243.20 | 1186.09 |
| $C_{FB}$ b=100 | 10.82 | 33.98 | 79.85 | 149.53 | 271.87 | 493.77 | 683.94 | 849.56 | 988.16 | 1243.20 | 1186.09 |
| $C_{FB}$ b=10000 | 11.91 | 34.03 | 79.85 | 149.53 | 271.87 | 493.77 | 683.94 | 849.56 | 988.16 | 1243.20 | 1186.09 |

Table 2.2: Robustness of $C_{FB}$ to $b$: average losses of $C_{CML}$, $C_{MML}$ and $C_{FB}$. m=500, n=p=1000, c=5, $\alpha = 1$ and $wa = wb = 1$.

changes from 0 to 10000. When the models are parsimonious, $SS_\gamma$ should be relatively small and becomes less dominant. Hence, when $b$ is large, we see a little influence from $b$ on the performance of $C_{FB}$: large $b$ favors small model, since the larger the b, the larger the $F_\gamma$.

We also tested the performance for other combinations of $\alpha$ and $b$, e.g., large $\alpha$ and small $b$, or small $\alpha$ and large $b$, etc.. We found that for all the values of $b$, $C_{FB}$ achieved better performance with small $\alpha$ than with large $\alpha$. Therefore, we set $\alpha = 1$ and $b = 0$ for other simulations.

Simulation 3: Investigate the robustness of $C_{FB}$ to the choice of $wa$ and $wb$.

Again, the data were generated using the same procedure. $m$, $n$, $p$, $c$ and $q$ are the same. $\alpha = 1$ and $b = 0$. The $C_{FB}$ was repeated on the same data for various $wa$ and $wb$: $(wa, wb)$ = (1,1), (0.1,0.1), (100,0.1), (0.1,100) and (100,100). $C_{CML}$ and $C_{MML}$ were again included for comparison. These combinations correspond to the five typical *Beta* densities: (1, 1) corresponds to the uniform density; (0.1, 0.1), a U-shaped density; (100,0.1), a increasing density; (0.1,100), a decreasing density and (100, 100), a symmetric unimodal density.

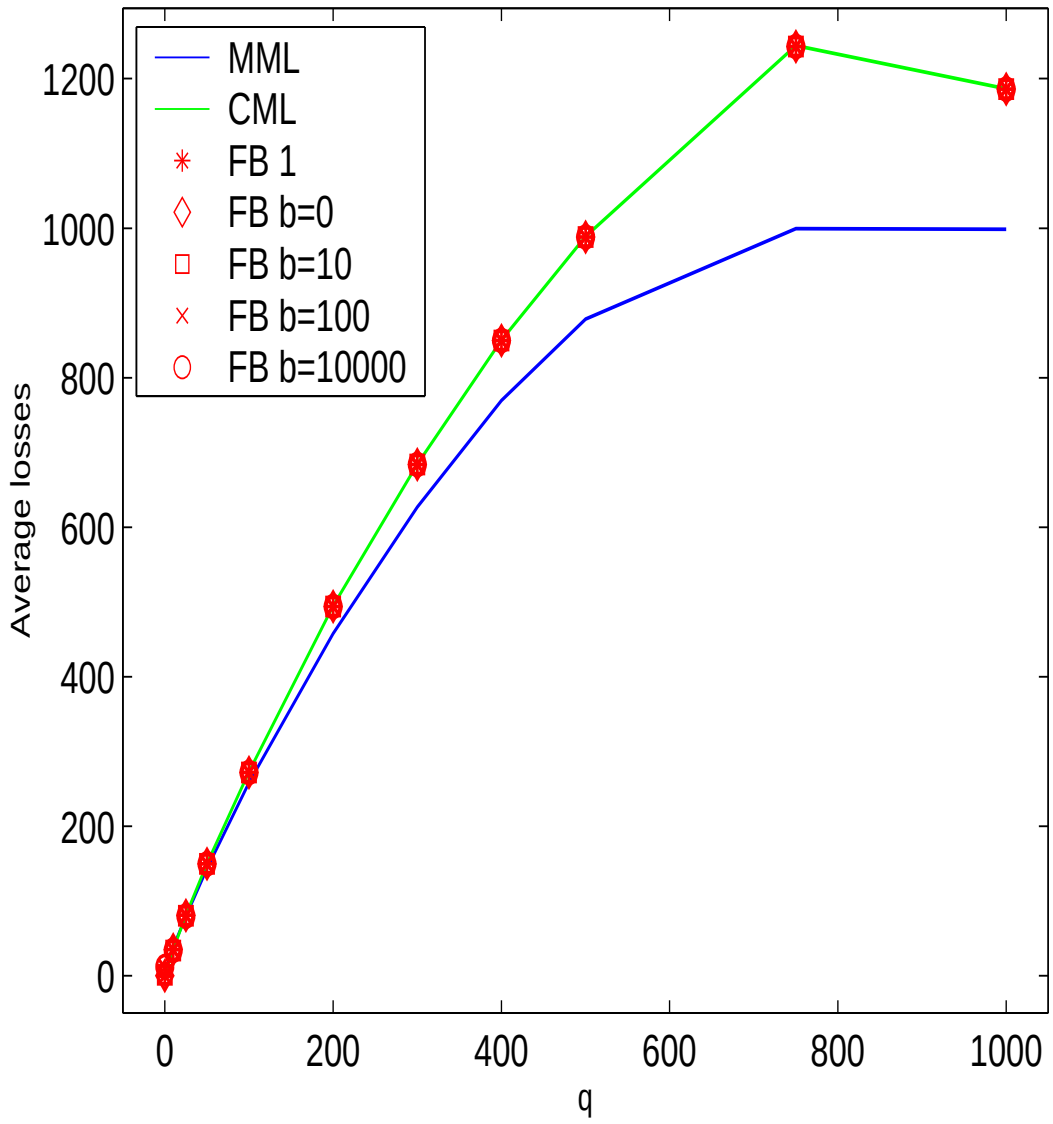


Figure 2.2: Robustness of $C_{FB}$ to $b$: average losses of $C_{CML}$, $C_{MML}$ and $C_{FB}$. m=500, n=p=1000, c=5, $\alpha = 1$ and $wa = wb = 1$.

For each fixed $\gamma$, $p(\gamma|\omega)$ is a unimodal function of $\omega$. $Beta(100, 100)$ is also unimodal and favors mid-sized models over others. Therefore, with prior $Beta(100, 100)$, $C_{FB}$ achieved better performance for models of moderately large (or moderately small) size. For example, in Table (2.3), we can see that when $q$ is between 100 and 750, $C_{FB}$ with $Beta(100, 100)$ is much better than others including $C_{CML}$, and sometimes it's even better than $C_{MML}$. But, towards the two ends, $C_{FB}$ is very poor. $Beta(100, 0.1)$ is increasing in $\omega$, which favors large models. Table (2.3) shows that the corresponding $C_{FB}$ is better when the model has 100 or more variables, but is pretty bad when the model is parsimonious. $Beta(0.1, 100)$ is just the opposite. The $C_{FB}$ under uniform and the $C_{FB}$ under $Beta(0.1, 0.1)$ are very simular. This may be because that $Beta(0.1, 0.1)$ is U-shaped and somehow levels off $p(\gamma|\omega)$ when $\omega$ is marginalized out. The changes of the performance of $C_{FB}$ for various $Beta$ densities can be more easily seen from Figure (2.3) for $q$ larger than 200.

The result shows that none of these $Beta$ priors yields better overall performance of $C_{FB}$ than that of $C_{CML}$ and $C_{MML}$. Actually, since all $Beta$ densities imply preference to models of certain sizes, uniform prior should be a more reasonable choice for $\omega$, unless there is a reason to choose other $Beta$ densities.

Based on all the results above, we set $\alpha = 1$, $b = 0$ and $wa = wb = 1$ in the rest simulations in this study.

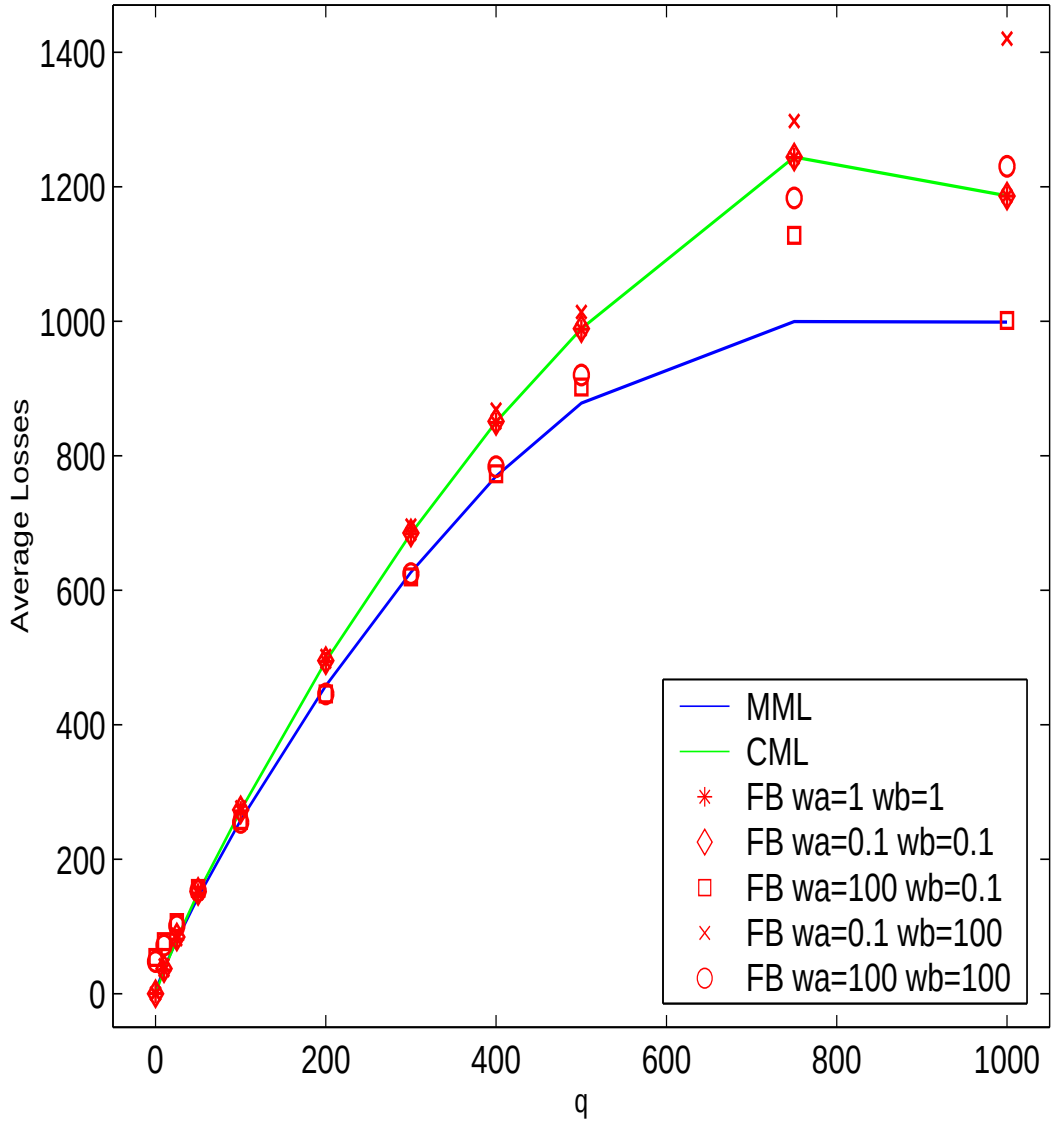


Figure 2.3: Robustness of $C_{FB}$ to $wa$ and $wb$: average losses of $C_{CML}$, $C_{MML}$ and $C_{FB}$. m=500, n=p=1000, c=5, $\alpha = 1$ and $b = 0$.

| q | 0 | 10 | 25 | 50 | 100 | 200 | 300 | 400 | 500 | 750 | 1000 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $C_{MML}$ | 3.84 | 38.07 | 79.41 | 143.50 | 258.09 | 457.84 | 626.73 | 769.69 | 878.41 | 999.60 | 998.70 |
| $C_{CML}$ | 0.17 | 34.79 | 81.04 | 150.11 | 272.58 | 494.60 | 684.49 | 850.34 | 988.84 | 1244.14 | 1186.44 |
| $C_{FB}$ wa=1 wb=1 | 0.13 | 35.09 | 80.92 | 149.66 | 271.87 | 493.77 | 683.94 | 849.56 | 988.16 | 1243.20 | 1186.09 |
| $C_{FB}$ wa=0.1 wb=0.1 | 0.00 | 37.23 | 84.13 | 151.90 | 273.48 | 495.24 | 685.01 | 851.00 | 989.06 | 1244.33 | 1186.15 |
| $C_{FB}$ wa=100 wb=0.1 | 53.91 | 77.24 | 105.69 | 156.66 | 257.52 | 445.60 | 619.81 | 773.16 | 902.21 | 1127.86 | 1001.13 |
| $C_{FB}$ wa=0.1 wb=100 | 0.00 | 42.08 | 87.49 | 154.33 | 277.08 | 502.20 | 695.93 | 868.60 | 1013.86 | 1297.84 | 1420.15 |
| $C_{FB}$ wa=100 wb=100 | 47.63 | 71.92 | 100.91 | 153.37 | 254.75 | 445.91 | 624.64 | 783.53 | 920.08 | 1183.28 | 1230.29 |

Table 2.3: Robustness of $C_{FB}$to $wa$ and $wb$: average losses of $C_{CML}$, $C_{MML}$ and $C_{FB}$. m=500, n=p=1000, c=5, $\alpha = 1$ and $b = 0$.

### 2.4.2 Compare Empirical Bayes with Fully Bayes via simulations

In this section, we compare the EB criteria with the FB criterion for two cases: $c = 5$ and $c = 25$. The data were generated in the same way as before. For each case, m=500, n=p=1000, $\alpha = 1$, $b = 0$ and $wa = wb = 1$. From the simulations in section 2.4.1, we can see that $C_{FB}$ and $C_{CML}$ are about the same (overall, $C_{FB}$ is slightly better than $C_{CML}$) , but $C_{FB}$ has never been as good as $C_{MML}$. In all those simulations, we chose $c = 5$. The hyper parameter, $c$, controls the size of the coefficients, $\beta$. When $c = 5$, the standard deviation of each component of $\beta$ is about 2.236 times of the standard deviation of the random noise. The signal is not very much stronger than the noise. In this case, we can see that $C_{MML}$ has done a much better job than both $C_{CML}$ and $C_{FB}$. When $c$ gets larger, the signal becomes stronger, and we expect that the difference between $C_{MML}$ and $C_{CML}$ and $C_{FB}$ becomes smaller. When $c$ is large enough, the signal will be so strong that all the criteria can be the same and all achieve very good performance. In Table (2.4), the results for both cases when $c = 5$ and $c = 25$ are presented. It can be easily seen that the three criteria differ much less when $c = 25$ than they do when $c = 5$. The difference in the performance between the two cases can be seen more apparently from

Average losses when $c = 5$

| q | 0 | 10 | 25 | 50 | 100 | 200 | 300 | 400 | 500 | 750 | 1000 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $C_{MML}$ | 3.84 | 38.07 | 79.41 | 143.50 | 258.09 | 457.84 | 626.73 | 769.69 | 878.41 | 999.60 | 998.70 |
| $C_{CML}$ | 0.17 | 34.79 | 81.04 | 150.11 | 272.58 | 494.60 | 684.49 | 850.34 | 988.84 | 1244.14 | 1186.44 |
| $C_{FB}$ | 0.13 | 35.09 | 80.92 | 149.66 | 271.87 | 493.77 | 683.94 | 849.56 | 988.16 | 1243.20 | 1186.09 |

Average losses when $c = 25$

| q | 0 | 10 | 25 | 50 | 100 | 200 | 300 | 400 | 500 | 750 | 1000 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $C_{MML}$ | 1.46 | 34.44 | 75.38 | 133.61 | 241.97 | 422.53 | 572.86 | 705.42 | 815.09 | 984.49 | 1003.84 |
| $C_{CML}$ | 0.25 | 34.85 | 78.15 | 137.47 | 250.50 | 436.56 | 592.76 | 727.47 | 848.53 | 1030.90 | 1003.84 |
| $C_{FB}$ | 0.13 | 34.46 | 77.77 | 137.37 | 250.04 | 436.27 | 592.36 | 727.21 | 848.40 | 1030.85 | 1003.84 |

Table 2.4: EB vs FB via simulations: Average losses for Bayes Least-Squares Procedures (BLS). m=500, n=p=1000, $\alpha = 1$, $b = 0$ and $wa = wb = 1$.

figure (2.4).

### 2.4.3 Bimodality in Fully Bayes posterior

Recall the discussion in section (2.3) in which we pointed out that the FB posterior (or $C_{FB}$) is bimodal. Figure (2.5) displays the plots of the log posteriors from the Fully Bayes procedure for different actual models from one replication of the data. For each actual model, the Fully Bayes log posterior of the best model of each size were calculated and plotted against the $p + 1$ distinct sizes. Here, "best" means that the model has the highest posterior probability among all the models of the same size. These plots show that the posterior has two modes: one at the null model or small model and the other one at the full model. Actually, for most of the models, the posterior is maximized at the full model.

Although the plots are generated from only one replication of the data, they do have the generality. For illustration purpose, the same plots but for $m = 50$ replications are presented in figure (2.6)

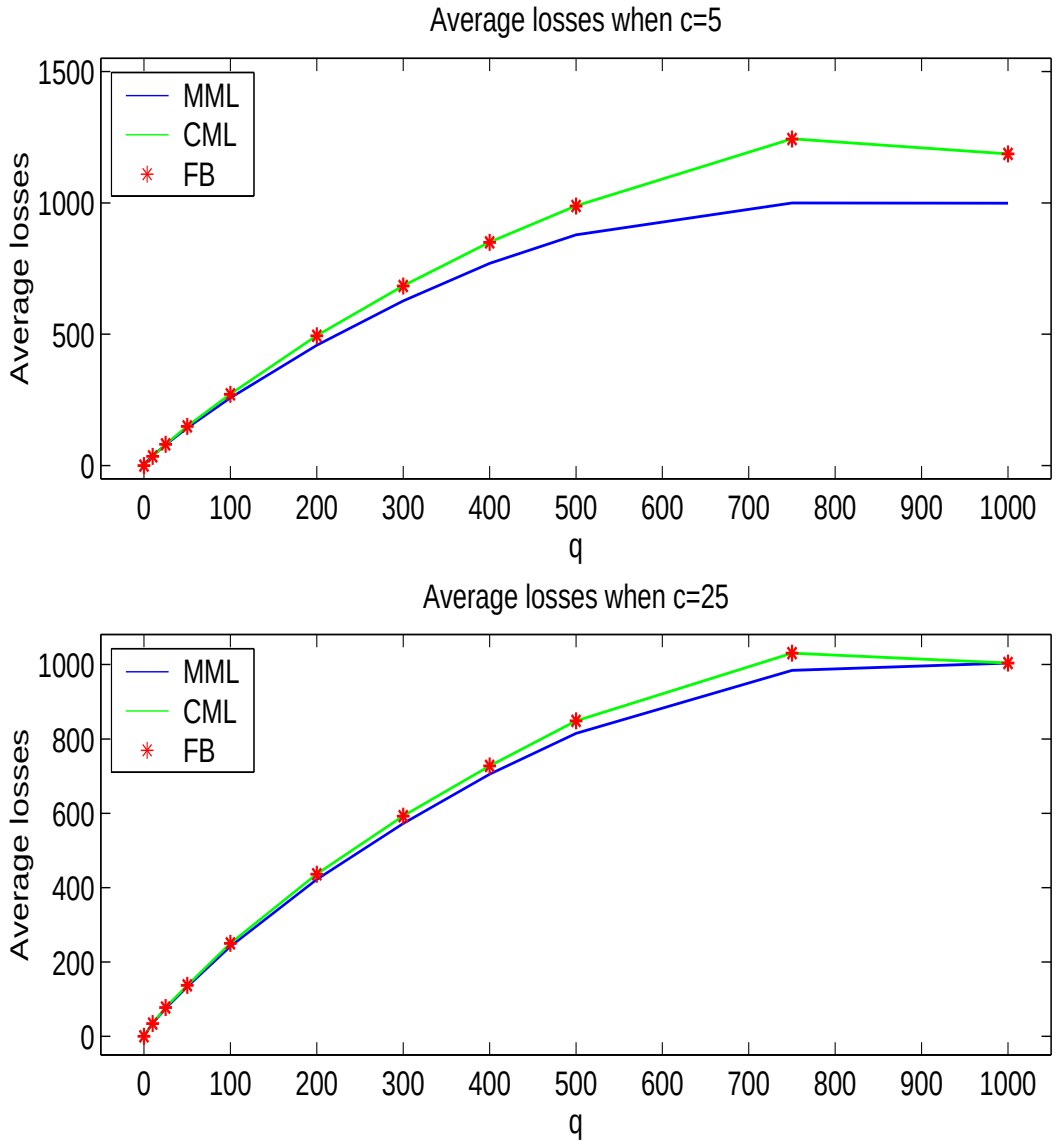


Figure 2.4: EB vs FB via simulations: Average losses for Bayes Least-Squares Procedures (BLS) m=500, n=p=1000, $\alpha = 1$, $b = 0$ and $wa = wb = 1$.

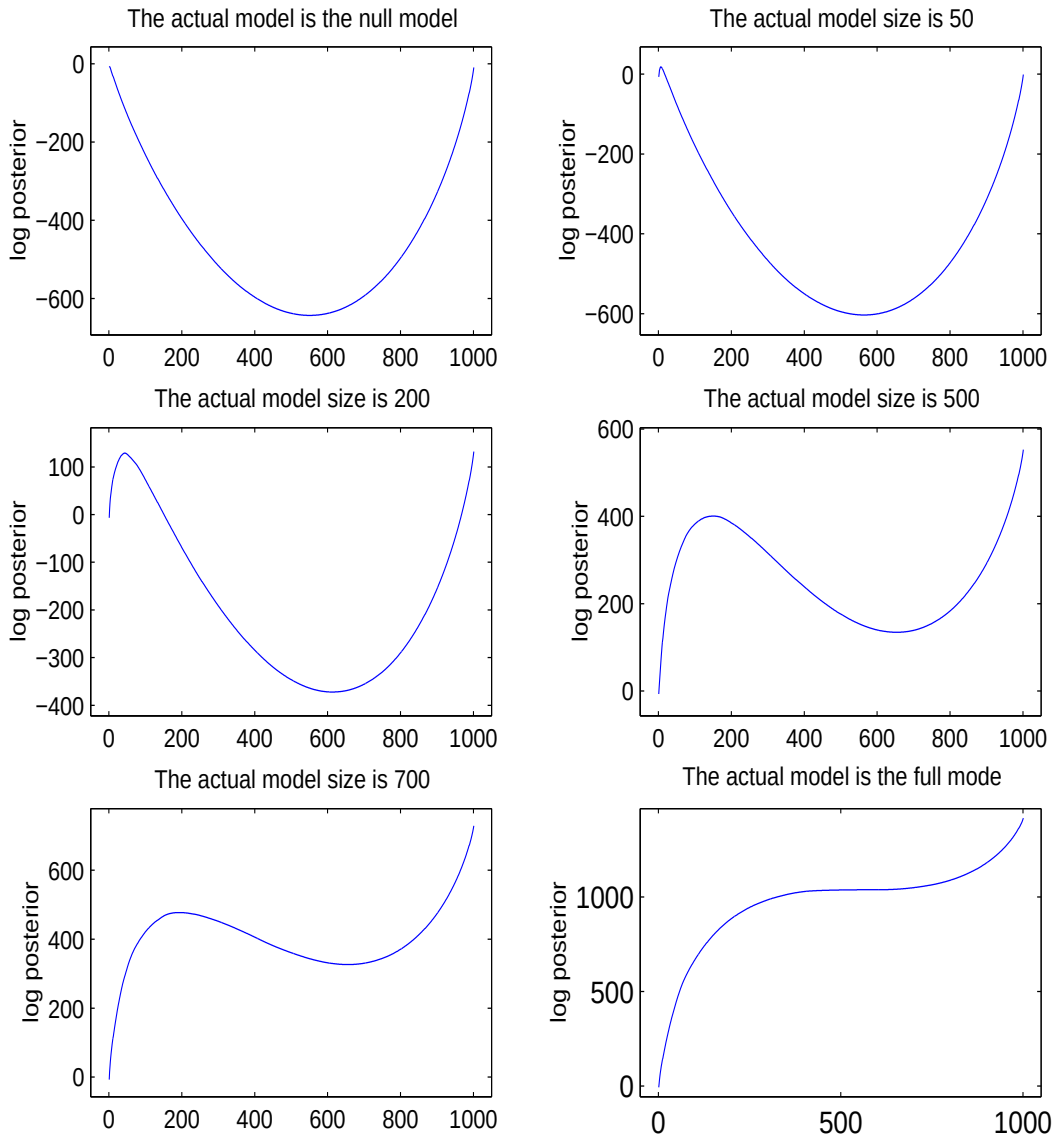


Figure 2.5: Bimodality 1: FB log posteriors from one replication of data. m=1, n=p=1000, $c = 5$, $\alpha = 1$, $b = 0$ and $wa = wb = 1$.

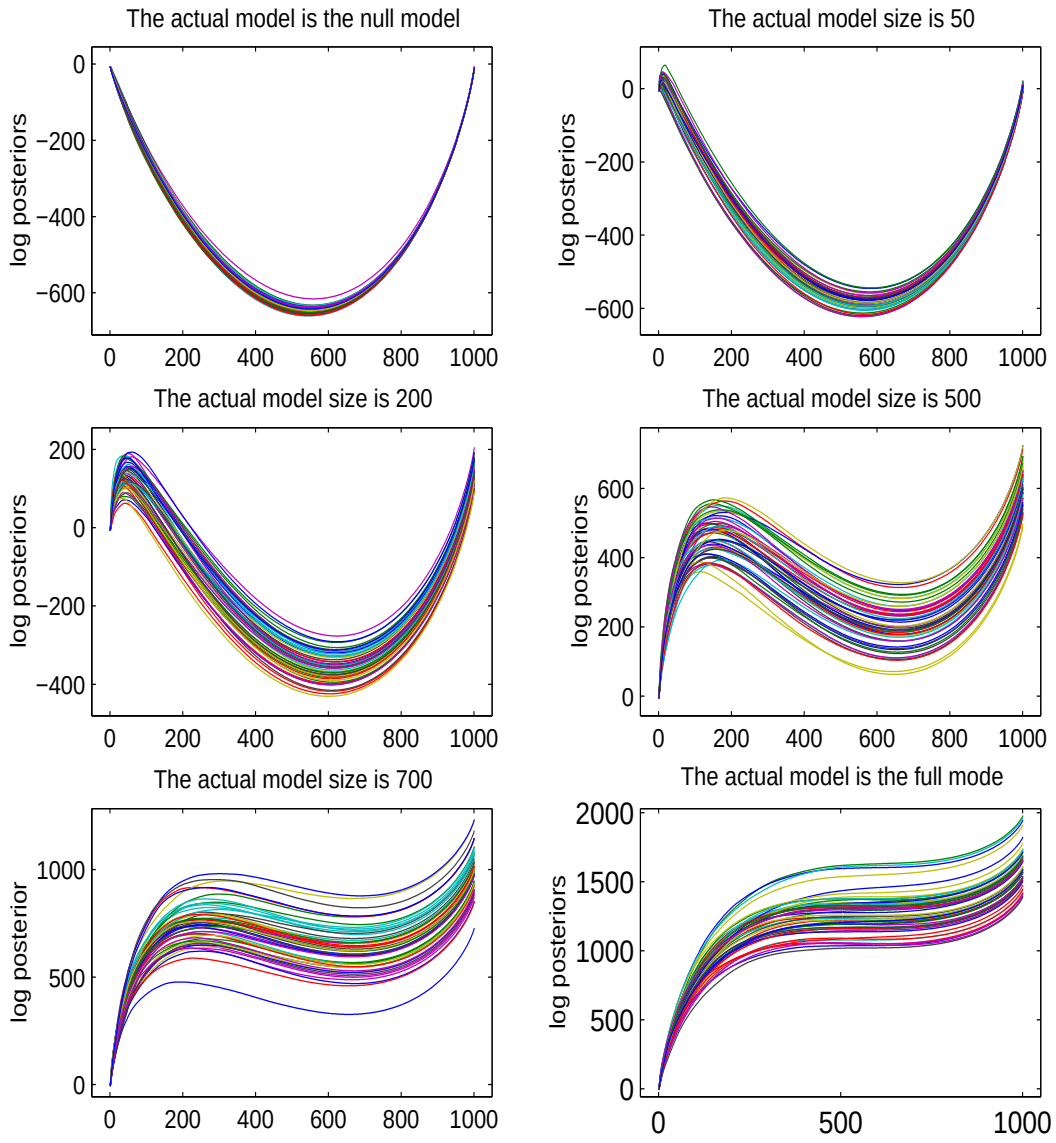


Figure 2.6: Bimodality 2: FB log posteriors from 50 replications of data. m=50, n=p=1000, $c = 5$, $\alpha = 1$, $b = 0$ and $wa = wb = 1$.

# Chapter 3

# Bayes Posterior Mean Procedures

After selection, the coefficients, $\beta_\gamma$, were estimated by the least-squares estimators in Chapter 2. Such estimators simply ignore the selection error and can lead to overestimation. Actually, a better estimator is the posterior mean of $\beta_\gamma$, since it's an admissible estimator of $\beta_\gamma$ under squared error loss.

Under the EB framework, the estimator is

$$\hat{\beta}_\gamma^{EB} = E(\beta_\gamma \mid Y, \gamma, \hat{c})$$

where $\hat{c}$ is EB estimator of $c$. Under the FB framework, the estimator is

$$\hat{\beta}_\gamma^{FB} = E(\beta_\gamma \mid Y, \gamma).$$

George and Foster (2000) demonstrated that the EB posterior mean estimator achieved better performance over the least-squares estimator. In section (3.1), we review the EB posterior mean estimator. In section (3.2), we derive a FB posterior mean estimator and in section (3.3), we compare the two and demonstrate the similarity between the $C_{CML}$ posterior mean estimator and the FB posterior mean estimator. The simulation results are reported in section (3.4).

## 3.1 Empirical Bayes criteria

Given the priors (2.3) and (2.4), the posterior mean, conditioning on $\gamma$, is

$$E(\beta_\gamma \mid Y, \gamma, \hat{c}) \;=\; \frac{c}{1+c}\hat{\beta}_\gamma \;=\; \left(1 - \frac{1}{1+c}\right)\hat{\beta}_\gamma, \tag{3.1}$$

where $\hat{\beta}_\gamma = (X_\gamma{}'X_\gamma)^{-1}X_\gamma{}'Y$ is the regular least-squares estimator. Once a model is selected by an EB criterion, $C_{CML}$ or $C_{MML}$, the corresponding posterior mean estimators can be obtained by substituting $c$ with $C_{CML}$ or $C_{MML}$ estimates of $c$ respectively. If $C_{CML}$ criterion is applied, the posterior mean is

$$\hat{\beta}_\gamma^{CML} = \left(1 - \frac{\sigma^2 q_\gamma}{SS_\gamma}\right)_+ \hat{\beta}_\gamma, \tag{3.2}$$

If $C_{MML}$ is applied, the posterior mean is

$$\hat{\beta}_\gamma^{MML} = \left(1 - \frac{1}{1+\hat{c}}\right)\hat{\beta}_\gamma, \tag{3.3}$$

where $\hat{c}$ is the maximum marginal likelihood estimator of $c$, which can be numerically computed. As we can see from (3.2) and (3.3), the posterior mean estimator is the least-squares estimator multiplied by a correction factor, which shrinks $\hat{\beta}_\gamma$ towards zeros. Such shrinkage should be more effective in improving the performance when the actual value of $c$ is small, since in such cases, it's even harder to distinguish between signal and noise.

## 3.2 Fully Bayes criteria

Instead of estimating $c$, FB estimates $\beta$ also by the conditional posterior mean, $E(\beta_\gamma \mid Y, \gamma)$, but with $c$ being integrated out with respect to the prior

(2.17). One way to compute $E(\beta_\gamma|Y,\gamma)$ is first to compute $p(\beta_\gamma|\hat{\beta}_\gamma,\gamma)$ by integrating out $c$ in the joint density of $\beta$ and $c$, $p(\beta_\gamma,c|\hat{\beta}_\gamma,\gamma)=p(\beta_\gamma|\hat{\beta}_\gamma,\gamma,c)\pi(c)$, and then compute the expectation. But, integrating out $c$ from the joint density can be difficult.

Another way to do that is first to compute $\pi(c \mid \hat{\beta}_\gamma, r)$ and then to compute the expectation of $E(\beta_\gamma \mid \hat{\beta}_\gamma, \gamma, c)$ with respect to $\pi(c \mid \hat{\beta}_\gamma, r)$, since

$$E(\beta_\gamma \mid \hat{\beta}_\gamma,\gamma) = E(E(\beta_\gamma \mid \hat{\beta}_\gamma,\gamma,c)) = \int_c E(\beta_\gamma \mid \hat{\beta}_\gamma,\gamma,c)\pi(c \mid \hat{\beta}_\gamma,\gamma)dc,$$

where $E(\beta_\gamma \mid \hat{\beta}_\gamma,\gamma,c) = \frac{c}{1+c}\hat{\beta}_\gamma$ and

$$\begin{aligned}\pi(c \mid \hat{\beta}_\gamma, r) &= \frac{p(\hat{\beta}_\gamma, c \mid r)}{p(\hat{\beta}_\gamma \mid r)} \\ &= \frac{p(\hat{\beta}_\gamma \mid r, c)\pi(c)}{\int_c p(\hat{\beta}_\gamma \mid r, c)\pi(c)dc}\end{aligned}$$

**Theorem 3.2.1.** *With $\pi(c)$ as (2.17) and $p(\hat{\beta}_\gamma \mid \gamma, c)$ as (2.15), the FB conditional posterior mean of $\beta_\gamma$ is*

$$\hat{\beta}_\gamma^{FB} = E(\beta_\gamma \mid \hat{\beta}_\gamma,\gamma) = \left\{1 - \frac{D(\frac{q_\gamma}{2}+\alpha+1)}{D(\frac{q_\gamma}{2}+\alpha)}\right\}\hat{\beta}_\gamma, \tag{3.4}$$

*where*

$$D(\frac{q_\gamma}{2}+\alpha) = \int_{t>1}^{\infty} t^{-(\frac{q_\gamma}{2}+\alpha)-1}\exp\left\{-\frac{SS_\gamma/2\sigma^2+b}{t}\right\}dt \tag{3.5}$$

*and*

$$D(\frac{q_\gamma}{2}+\alpha+1) = \int_{t>1}^{\infty} t^{-(\frac{q_\gamma}{2}+\alpha+1)-1}\exp\left\{-\frac{SS_\gamma/2\sigma^2+b}{t}\right\}dt \tag{3.6}$$

Proof: Given $\pi(c)$ in (2.17) and $p(\hat{\beta}_\gamma \mid \gamma, c)$ in (2.15), The joint density of $\hat{\beta}_\gamma$ and $c$ is

$$\begin{aligned}
p(\hat{\beta}_\gamma, c \mid r) &= p(\hat{\beta}_\gamma \mid r, c)\pi(c) \\
&= (2\pi)^{-q_\gamma/2}|(1+c)(X_\gamma{}'X_\gamma)^{-1}\sigma^2|^{-1/2}e^{-\frac{\hat{\beta}_\gamma'(X_\gamma{}'X_\gamma)\hat{\beta}_\gamma}{2(1+c)\sigma^2}} \\
&\quad \cdot M(1+c)^{-\alpha-1}e^{-\frac{b}{1+c}} \\
&= K_{q_\gamma}(1+c)^{-\frac{q_\gamma}{2}-\alpha-1}\exp\left\{-\frac{\hat{\beta}_\gamma'(X_\gamma{}'X_\gamma)\hat{\beta}_\gamma/2\sigma^2+b}{1+c}\right\} \\
&= K_{q_\gamma}(1+c)^{-\frac{q_\gamma}{2}-\alpha-1}\exp\left\{-\frac{SS_\gamma/2\sigma^2+b}{1+c}\right\},
\end{aligned}$$

where $K_{q_\gamma} = (2\pi)^{-q_\gamma/2}|(X_\gamma{}'X_\gamma)^{-1}\sigma^2|^{-1/2}M$ and $c > 0$.

$$\begin{aligned}
p(\hat{\beta}_\gamma \mid r) &= \int_c p(\hat{\beta}_\gamma, c \mid r)dc \\
&= K_{q_\gamma}\int_{c>0}^{\infty}(1+c)^{-\frac{q_\gamma}{2}-\alpha-1}\exp\left\{-\frac{SS_\gamma/2\sigma^2+b}{1+c}\right\}dc \\
&= K_{q_\gamma}\int_{t>1}^{\infty}t^{-\frac{q_\gamma}{2}-\alpha-1}\exp\left\{-\frac{SS_\gamma/2\sigma^2+b}{t}\right\}dt.
\end{aligned}$$

Then

$$\begin{aligned}
\pi(c \mid \hat{\beta}_\gamma, r) &= \frac{p(\hat{\beta}_\gamma, c \mid r)}{p(\hat{\beta}_\gamma \mid r)} \\
&= \frac{(1+c)^{-\frac{q_\gamma}{2}-\alpha-1}\exp\left\{-\frac{SS_\gamma/2\sigma^2+b}{1+c}\right\}}{\int_{t>1}^{\infty}t^{-\frac{q_\gamma}{2}-\alpha-1}\exp\left\{-\frac{SS_\gamma/2\sigma^2+b}{t}\right\}dt}.
\end{aligned}$$

Now,

$$\begin{aligned}
E(\beta_\gamma \mid \hat{\beta}_\gamma, \gamma) &= E(E(\beta_\gamma \mid \hat{\beta}_\gamma, \gamma, c)) \\
&= E^{c \mid \hat{\beta}_\gamma, \gamma}(\frac{c}{1+c}\hat{\beta}_\gamma) \\
&= \int_c \frac{c}{1+c}\hat{\beta}_\gamma \pi(c \mid \hat{\beta}_\gamma, \gamma) dc \\
&= \int_c (1 - \frac{1}{1+c})\hat{\beta}_\gamma \pi(c \mid \hat{\beta}_\gamma, \gamma) dc \\
&= \hat{\beta}_\gamma - \int_c \frac{1}{1+c}\hat{\beta}_\gamma \pi(c \mid \hat{\beta}_\gamma, \gamma) dc \\
&= \hat{\beta}_\gamma - \hat{\beta}_\gamma \frac{\int_{c>0}(1+c)^{-\frac{q_\gamma}{2}-\alpha-1-1} \exp\left\{-\frac{SS_\gamma/2\sigma^2+b}{1+c}\right\} dc}{\int_t t^{-\frac{q_\gamma}{2}-\alpha-1} \exp\left\{-\frac{SS_\gamma/2\sigma^2+b}{t}\right\} dt} \\
&= \hat{\beta}_\gamma - \hat{\beta}_\gamma \frac{\int_{t>1} t^{-\frac{q_\gamma}{2}-\alpha-1-1} \exp\left\{-\frac{SS_\gamma/2\sigma^2+b}{t}\right\} dt}{\int_t t^{-\frac{q_\gamma}{2}-\alpha-1} \exp\left\{-\frac{SS_\gamma/2\sigma^2+b}{t}\right\} dt} \\
&= \left\{1 - \frac{D(\frac{q_\gamma}{2}+\alpha+1)}{D(\frac{q_\gamma}{2}+\alpha)}\right\} \hat{\beta}_\gamma, \qquad (3.7)
\end{aligned}$$

where

$$D(\frac{q_\gamma}{2}+\alpha) = \int_{t>1}^{\infty} t^{-(\frac{q_\gamma}{2}+\alpha)-1} \exp\left\{-\frac{SS_\gamma/2\sigma^2+b}{t}\right\} dt$$

and

$$D(\frac{q_\gamma}{2}+\alpha+1) = \int_{t>1}^{\infty} t^{-(\frac{q_\gamma}{2}+\alpha+1)-1} \exp\left\{-\frac{SS_\gamma/2\sigma^2+b}{t}\right\} dt.$$

Similar to EB posterior mean estimators, the FB posterior mean estimator is also the least-squares estimator multiplied by a correction factor, which shrinks it towards zero.

## 3.3 Empirical Bayes vs. Fully Bayes

In Chapter 2, we have shown the similarity between $C_{FB}$ and $C_{CML}$ when $\beta_\gamma$ are estimated by the least-squares estimator after selection. Next, we will show that the posterior mean estimators of $\beta_\gamma$ following the two selection procedures are also very close.

It can be easily seen that

$$D(\frac{q_\gamma}{2}+\alpha+1)=\frac{\int_0^{\frac{SS_\gamma}{2\sigma^2+b}} t^{\frac{q_\gamma}{2}+\alpha+1-1}e^{-t}dt}{(\frac{SS_\gamma}{2\sigma^2}+b)^{q_\gamma/2+\alpha+1}}$$

and

$$D(\frac{q_\gamma}{2}+\alpha)=\frac{\int_0^{\frac{SS_\gamma}{2\sigma^2+b}} t^{\frac{q_\gamma}{2}+\alpha-1}e^{-t}dt}{(\frac{SS_\gamma}{2\sigma^2}+b)^{q_\gamma/2+\alpha}}.$$

Let

$$G(\frac{q_\gamma}{2}+\alpha)=\int_0^{\frac{SS_\gamma}{2\sigma^2+b}} t^{\frac{q_\gamma}{2}+\alpha-1}e^{-t}dt,$$

then

$$D(\frac{q_\gamma}{2}+\alpha+1)=\frac{G(\frac{q_\gamma}{2}+\alpha+1)}{(\frac{SS_\gamma}{2\sigma^2}+b)^{q_\gamma/2+\alpha+1}}$$

and

$$D(\frac{q_\gamma}{2}+\alpha)=\frac{(\frac{SS_\gamma}{2\sigma^2}+b)^{\frac{q_\gamma}{2}+\alpha}e^{-(\frac{SS_\gamma}{2\sigma^2}+b)}+G(\frac{q_\gamma}{2}+\alpha+1)}{(\frac{q_\gamma}{2}+\alpha)(\frac{SS_\gamma}{2\sigma^2}+b)^{q_\gamma/2+\alpha}}$$

The first term in the numerator,

$$(\frac{SS_\gamma}{2\sigma^2}+b)^{\frac{q_\gamma}{2}+\alpha}e^{-(\frac{SS_\gamma}{2\sigma^2}+b)}$$

| q | 0 | 10 | 25 | 50 | 100 | 200 | 300 | 400 | 500 | 750 | 1000 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| MMLLS | 2.95 | 36.59 | 77.94 | 143.18 | 258.99 | 460.34 | 622.40 | 768.10 | 878.64 | 999.00 | 1000.62 |
| CMLLS | 0.08 | 34.83 | 79.73 | 150.14 | 273.49 | 493.95 | 677.41 | 847.68 | 985.80 | 1243.62 | 1169.12 |
| FBLS | 0.04 | 35.00 | 79.90 | 149.37 | 273.01 | 492.98 | 676.64 | 847.07 | 985.30 | 1242.88 | 1167.91 |
| MMLPM | 1.00 | 31.42 | 68.98 | 127.75 | 228.12 | 395.31 | 524.55 | 633.55 | 712.97 | 798.04 | 836.90 |
| CMLPM | 0.08 | 34.52 | 78.87 | 147.92 | 268.19 | 480.67 | 654.60 | 813.87 | 939.78 | 1163.34 | 1030.00 |
| FBPM | 0.03 | 34.46 | 78.61 | 146.58 | 267.11 | 479.09 | 653.26 | 812.70 | 938.71 | 1162.12 | 1028.50 |

Table 3.1: EB vs FB via simulations: Average losses for BLS and Bayes Posterior Mean (BPM) Procedures. m=500, n=p=1000, c=5, $\alpha = 1$, $b = 0$ and $wa = wb = 1$.

is actually a rough approximation to $G(\frac{q_\gamma}{2} + \alpha + 1)$. Therefore, the numerator is approximately $2\,G(\frac{q_\gamma}{2} + \alpha + 1)$ and

$$\hat{\beta}_\gamma^{FB} \approx \left\{1 - \frac{\sigma^2(q_\gamma + 2\alpha)}{SS_\gamma + b}\right\}_+ \hat{\beta}_\gamma. \tag{3.8}$$

When $\alpha$ and $b$ are small, $\hat{\beta}_\gamma^{FB}$ is approximately the same as

$$\hat{\beta}_\gamma^{CML} = \left(1 - \frac{\sigma^2 q_\gamma}{SS_\gamma}\right)_+ \hat{\beta}_\gamma,$$

## 3.4 Simulations

The data were generated in the same way as it was in (2.4). $m$=500, $n = p = 1000$, $c = 5$, $\alpha = 1$, $b$=1 and $wa = wb = 1$. 500 replications were generated for each set of $q = 0$, 10, 25, 50, 100, 200, 300, 400, 500, 750, and 1000. For each set of data generated, $C_{CML}$, $C_{MML}$ and $C_{FB}$ were applied to select the model, and then the coefficients of the model selected were estimated by corresponding posterior mean estimators. In Table 3.1 and Figure 3.1, "CMLLS" stands for the procedure in which the $C_{CML}$ selection criterion was applied and $\beta_\gamma$ was estimated by the least-squares estimator. Similarly, "CMLPM" stands for the procedure in which the $C_{CML}$ selection

criterion was applied and $\beta_\gamma$ was estimated by the posterior mean estimator. The others have the similar meanings. The predictive losses were computed for all the six procedures and were averaged over the 500 replications.

The average losses were presented in Table 3.1 and were plotted in Figure 3.1. The simulation results show that

**1** all the three procedures can achieve better performance with posterior mean estimators than with least-squares estimators;

**2** $C_{MML}$ benefits more from estimating $\beta_\gamma$ by the posterior mean than $C_{CML}$ and $C_{FB}$ do;

**3** $C_{CML}$ and $C_{FB}$ are again very close and $C_{FB}$ is slightly uniformly better than $C_{CML}$.

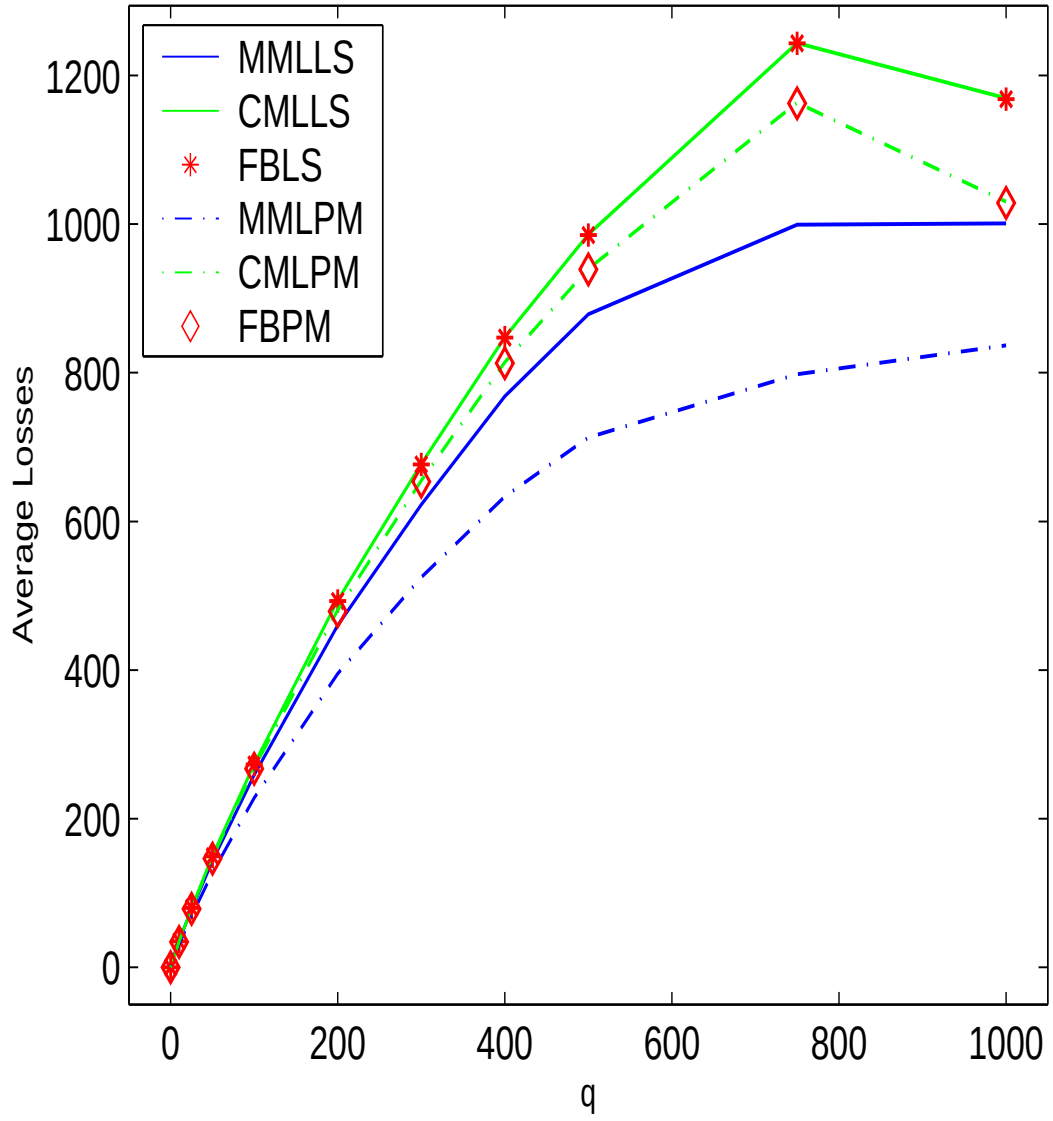


Figure 3.1: EB vs FB via simulations: Average losses for BLS and Bayes Posterior Mean (BPM) Procedures. m=500, n=p=1000, c=5, $\alpha = 1$, $b = 0$ and $wa = wb = 1$.

# Chapter 4

# Model Averaging Procedures

The results reported in Chapter 2 and 3 have shown that both $C_{FB}$ and $C_{CML}$ cannot capture the model information contained in the data as effectively as $C_{MML}$ does. Actually, the estimators of $c$ and $\omega$ in $C_{MML}$ are the posterior modes under the uniform priors on $c$ and $\omega$. These estimators successfully incorporate the information about the hyperparameters contained in all the possible models, and consequently benefit the selection procedure. $C_{FB}$ and $C_{CML}$ are obtained through a conditional procedure in the sense that when $c$ and $\omega$ are estimated or marginalized out, $\gamma$ is treated as known. The consequence is that the information about the hyperparameters contained in other models has been simply ignored. Actually, given the fact that both large models with small coefficients and small models with large coefficients can produce the same data, it may not be very surprising that $C_{FB}$ and $C_{CML}$ won't be able to distinguish the two kinds of models.

As an alternative, Bayesian model averaging may overcome the shortcoming in $C_{FB}$ and $C_{CML}$. Instead of estimating the model($\beta_\gamma$) after selection, the model averaging procedures estimate the model by a posterior mean of $\beta$ that is a weighted average of $E(\beta_\gamma|Y,\gamma)$ and the weights are the posterior

probabilities of the models, $\pi(\gamma|Y)$. Actually, Bayesian model averaging has been shown to be a better choice when the goal is exclusively prediction (see Raftery *et al.* 1997 and Clyde *et al.* 1998).

In this chapter, we will investigate the performance of the FB model averaging and compare the FB model averaging posterior mean with EB model averaging posterior mean, which is based on a multiple shrinkage estimator in the wavelet context (Clyde *et al.* 1998, Clyde and George,1999, 2000). In this chapter, we will limit our discussion to the case of orthogonal $X$.

## 4.1 Empirical Bayes Model Averaging

Under the EB model averaging framework, the posterior mean of $\beta$ is

$$E(\beta \mid Y, c, \omega) = \sum_{\gamma} E(\beta \mid Y, \gamma, c, \omega)\pi(\gamma \mid Y, c, \omega). \tag{4.1}$$

Here, directly computing $E(\beta \mid Y, c, \omega)$ is not desirable because 1) it requires averaging over all the possible models, which is impractical when $p$ is large, and 2) $\pi(\gamma \mid Y, c, \omega)$ does not have a closed form (or, it's only known up to a normalizing constant). To bypass the difficulties, one can apply a stochastic search that samples from the entire model space, and then compute the sample mean of $E(\beta \mid Y, \gamma, c, \omega)$ over the models sampled. Or, one can apply the SSVS (Stochastic Search Variable Selection) proposed in George and McCulloch (1993 and 1997) to identify a promising subset of models and then proceed the model averaging among this much smaller subset.

When $X$ is orthogonal, the EB model averaging can be much simplified

and is very straightforward. In this case, $\beta$ can be obtained basing on a multiple shrinkage estimator of $\beta$ in the context of wavelet regression (see Clyde et al., 1998, and Clyde and George, 1999 and 2000). Here is how the orthogonality helps: when $X$ is orthogonal, we can rewrite model (1.1) as

$$Y_i = \beta_i + \epsilon_i,$$

where $\beta_i$ is the $ith$ component of $\beta$ and $i = 1, 2, ..., p$. Let $\gamma^* = (\gamma_1, \gamma_2, ..., \gamma_p)$ be a vector of binary variables, in which $\gamma_i$ is a Bernoulli variable with $\pi(\gamma_1 = 1) = \omega$, and $\gamma_1 = 1$ stands for inclusion of the $i$th variable. Here, $\gamma^*$ is different from the index, $\gamma$, used before.

Given $X$ being orthogonal, it follows naturally from (2.3) that,

$$\beta_i \mid \gamma_i \ \sim \ N(0, c\,\gamma_i\,\sigma^2).$$

In addition,

$$\begin{aligned} \gamma_i \ &\sim \ Bernoulli(\omega), \\ \epsilon \ &\sim \ N(0, \sigma^2), \end{aligned}$$

and $\pi(c)$ and $\pi(\omega)$ are the same as (2.17) and (2.4). The posterior mean of $\beta_i$ is then

$$\begin{aligned} E(\beta_i \mid Y_i, c, \omega) &= E(E(\beta_i \mid Y_i, \gamma_i, c, \omega)) = E(\gamma_i \frac{c}{1+c}\hat{\beta}_i) \\ &= E(\gamma_i \mid Y_i, c, \omega)\frac{c}{1+c}\hat{\beta}_i \\ &= \pi(\gamma_i = 1 \mid Y_i, c, \omega)\frac{c}{1+c}\hat{\beta}_i. \end{aligned}$$

It can be shown that, given $Y_i$ and $c$, $\gamma_i$, $i = 1, 2, ..., p$ are independent Bernoulli variables with

$$\pi(\gamma_i = 1 \mid Y, c, \omega) = \frac{O_i}{1 + O_i} \tag{4.2}$$

where $O_i$ is the posterior odds that $\gamma_i = 1$ and

$$O_i = (1 + c)^{-1/2} \left( \frac{\omega}{1 - \omega} \right) \exp \left\{ \frac{1}{2\sigma^2} \frac{c}{1 + c} \hat{\beta}_i^2 \right\}. \tag{4.3}$$

Therefore,

$$E(\beta_i \mid Y_i, c, \omega) \quad = \quad \frac{O_i}{1 + O_i} \frac{c}{1 + c} \hat{\beta}_i. \tag{4.4}$$

By substituting $c$ and $\omega$ with the values chosen, we can obtain the corresponding model averaging estimators. For example, if we choose $c$ and $\omega$ such that the dimension penalty, $F(c, \omega)$=2, $\log n$ or $2 \log p$, we will obtain the corresponding $AIC/C_p$, $BIC$ and $RIC$ model averaging estimator. Clyde et al. (1998) chose $\omega = 0.5$ and the RIC criterion (i.e, $c$ was obtained by solving $F(c, \omega) = 2 \log p$), and showed that the model averaging estimator (a multiple shrinkage estimator of the wavelet regression coefficients) achieved excellent performance.

$C_{MML}$ estimators of $c$ and $\omega$ are obtained from the marginal likelihood function and can be easily put into (4.4) to substitute. But, $C_{CML}$ model averaging estimator can not be obtained similarly. The $C_{CML}$ estimators of $c$ and $\omega$ depend on $\gamma$ and hence prevent model averaging from being implemented with $C_{CML}$ criterion.

## 4.2 Fully Bayes Model Averaging

Similarly, the FB model averaging approach to the problem is to estimate $\beta$ by the posterior mean, but, where $c$ and $\omega$ are integrated out with respect to the hyperpriors (2.17) and (2.4). In this section, we will derive the FB model averaging estimator of $\beta$.

Directly computing $E(\beta_\gamma|Y,\gamma)$ can be difficult. But, things can be simplified by using conditional expectation:

$$\begin{aligned} E(\beta|Y) &= E(E(\beta_\gamma|Y,\gamma)) \\ &= \sum_{r=1}^{2^p} E(\beta_\gamma|Y,\gamma)\cdot\pi(\gamma|Y) \end{aligned} \tag{4.5}$$

Here, we have the same difficulties as what we had with the EB model averaging procedures: the conditional posterior means need to be averaged over all the models, and the posterior density function, $\pi(\gamma|Y)$, can only be known up to a normalizing constant. In section 4.1, we get around these difficulties by taking advantages of the fact that $X$ is orthogonal and $\beta_i$s are independent conditional on $c$ and $\omega$. However, FB model averaging cannot take advantages of this, due to the dependence of $\beta_i$ on $c$. A brief explanation is given next.

$$\begin{aligned} E(\beta_i \mid \hat{\beta}_i) &= E(E(\beta_i \mid \hat{\beta}_i, \gamma_i)) \\ &= E(\beta_i \mid \hat{\beta}_i, \gamma_i = 1)\,\pi(\gamma_i = 1|\hat{\beta}_i) + E(\beta_i \mid \hat{\beta}_i, \gamma_i = 0)\,\pi(\gamma_i = 0|\hat{\beta}_i) \\ &= E(\beta_i \mid \hat{\beta}_i, \gamma_i = 1)\,\pi(\gamma_i = 1|\hat{\beta}_i). \end{aligned}$$

$E(\beta_i \mid \hat{\beta}_i, \gamma_i = 1)$ can be computed in a similar fashion

$$\begin{aligned} E(\beta_i \mid \hat{\beta}_i, \gamma_i = 1) &= E(E(\beta_i \mid \hat{\beta}_i, \gamma_i = 1, c)) \\ &= \int_c E(\beta_i \mid \hat{\beta}_i, \gamma_i = 1, c)\pi(c|\hat{\beta}_i)dc. \end{aligned}$$

Since all $\beta_i$s depend on $c$, to obtain $\pi(c|\hat{\beta}_i)$ we need know the joint density of $(\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p)$ and then compute the marginal density of $\hat{\beta}_i$ out of it. Such a process can be very intractable. For this reason, we adopted (4.5).

**Theorem 4.2.1.** *Consider the variable selection problem for the linear model (1.1). Given the priors (2.3), (2.4), (2.17) and* Beta$(wa, wb)$ *on $\omega$, the FB model averaging estimator of $\beta$ is*

$$\begin{aligned} E(\beta|Y) &= \sum_\gamma E(\beta_\gamma|Y, \gamma)\, \pi(\gamma|Y) \\ &= \sum_\gamma \hat{\beta}_\gamma \left\{ 1 - \frac{D(\frac{q_\gamma}{2} + \alpha + 1)}{D(\frac{q_\gamma}{2} + \alpha)} \right\} K\, \frac{K_{q_\gamma}\, D(\frac{q_\gamma}{2} + \alpha)\, \pi(\gamma)\, e^{\frac{SS_\gamma}{2\sigma^2}}}{m(y)}, \end{aligned}$$

*where $D(\frac{q_\gamma}{2} + \alpha)$ and $D(\frac{q_\gamma}{2} + \alpha + 1)$ are the same as (3.5) and (3.6),*

$$K = (2\pi)^{-\frac{n}{2}} (\sigma^2)^{-\frac{n}{2}} e^{-\frac{Y'Y}{2\sigma^2}},$$

$$K_{q_\gamma} = (2\pi)^{-\frac{q_\gamma}{2}} |(X_\gamma{}' X_\gamma)^{-1}\sigma^2|^{-\frac{1}{2}} M,$$

*$\pi(\gamma)$ is the same as (2.22) and $m(y)$ is the marginal density of $Y$.*

Proof: It can be shown that $E(\beta_\gamma|Y, \gamma) = E(\beta_\gamma|\hat{\beta}_\gamma, \gamma)$ and the posterior of $\gamma$

$$\pi(\gamma|Y) = \frac{g(Y|\hat{\beta}_\gamma, \gamma)\, p(\hat{\beta}_\gamma \mid \gamma)\, \pi(\gamma)}{m(y)}.$$

where $g(Y|\hat{\beta}_\gamma, \gamma)$ is the same as (2.12), i.e.,

$$g(Y|\hat{\beta}_\gamma, \gamma) \;=\; (2\pi)^{-\frac{n}{2}}(\sigma^2)^{-\frac{n}{2}} e^{-\frac{Y'Y - SS_\gamma}{2\sigma^2}},$$

and the joint density of $\hat{\beta}_\gamma$ and $\gamma$

$$p(\hat{\beta}_\gamma|\gamma)\pi(\gamma) \;=\; \left(\int_c p(\hat{\beta}_\gamma|\gamma, c)\, \pi(c)\, dc\right) \pi(\gamma).$$

Now, given

$$\hat{\beta}_\gamma|r, c \sim N(0,\ (1+c)(X_\gamma{}'X_\gamma)^{-1}\sigma^2),$$

$$\pi(c) = M(1+c)^{-\alpha-1} e^{-\frac{b}{1+c}}, \quad c > 0$$

and $\pi(\gamma)$ in (2.22), we have

$$\begin{aligned}
\int_c p(\hat{\beta}_\gamma|r, c)\pi(c) &= K_{q_\gamma} \int_{c>0} (1+c)^{-\frac{q_\gamma}{2}-\alpha-1} \exp\left\{-\frac{\hat{\beta}'_\gamma (X_\gamma{}'X_\gamma)\hat{\beta}_\gamma/2\sigma^2 + b}{(1+c)}\right\} dc \\
&= K_{q_\gamma} \int_{t>1} t^{-\frac{q_\gamma}{2}-\alpha-1} \exp\left\{-\frac{SS_\gamma/2\sigma^2 + b}{t}\right\} dc \\
&= K_{q_\gamma} \cdot D(\frac{q_\gamma}{2} + \alpha).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\pi(\gamma|\hat{\beta}_\gamma) &= \frac{g(Y|\hat{\beta}_\gamma, \gamma)\; p(\hat{\beta}_\gamma \mid \gamma)\; \pi(\gamma)}{m(y)} \\
&= \frac{(2\pi)^{-\frac{n}{2}}(\sigma^2)^{-\frac{n}{2}} e^{-\frac{Y'Y - SS_\gamma}{2\sigma^2}}\; K_{q_\gamma} \cdot D(\frac{q_\gamma}{2} + \alpha)\; \pi(\gamma)}{m(y)} \\
&= K \;\cdot\; \frac{K_{q_\gamma} \cdot D(\frac{q_\gamma}{2} + \alpha)\; \pi(\gamma)\; e^{\frac{SS_\gamma}{2\sigma^2}}}{m(y)}.
\end{aligned}$$

In Chapter 3, it has been shown that

$$E(\beta_\gamma|\hat{\beta}_\gamma,\gamma) \;=\; \hat{\beta}_\gamma\left\{1-\frac{D(\frac{q_\gamma}{2}+\alpha+1)}{D(\frac{q_\gamma}{2}+\alpha)}\right\}.$$

Therefore, the posterior mean of $\beta$ is

$$\begin{aligned} E(\beta|Y) &= \sum_\gamma E(\beta_\gamma|Y,\gamma)\,\pi(\gamma|Y) \\ &= \sum_\gamma \hat{\beta}_\gamma\left\{1-\frac{D(\frac{q_\gamma}{2}+\alpha+1)}{D(\frac{q_\gamma}{2}+\alpha)}\right\}\,K\,\frac{K_{q_\gamma}\,D(\frac{q_\gamma}{2}+\alpha)\,\pi(\gamma)\,e^{\frac{SS_\gamma}{2\sigma^2}}}{m(y)}. \end{aligned}$$

When $\gamma=1$, $E(\beta_\gamma|Y,\gamma=1)=0$ and

$$\begin{aligned} \pi(\gamma|Y) &= \frac{p(Y|\gamma=1)\,\pi(\gamma=1)}{m(y)} \\ &= \frac{(2\pi)^{-\frac{n}{2}}(\sigma^2)^{-\frac{n}{2}}e^{-\frac{Y'Y}{\sigma^2}}\pi(\gamma=1)}{m(y)} \\ &= \frac{K\,\pi(\gamma=1)}{m(y)} \end{aligned}$$

## 4.3 Empirical Bayes vs. Fully Bayes

As we can see from the $C_{MML}$ model averaging estimator (4.4), the conditional posterior mean of $\beta_i$ has been shrunk by the posterior inclusion probability of the $i$th variable. Such shrinkage further denoises the data and yields a simpler model. The shrinkage is a very nice property and it helps to weaken weak-signal variables and make strong-signal variables stand out.

Since the $C_{CML}$ estimators of $c$ and $\omega$ are conditional on $\gamma$, $C_{CML}$ can't reap the benefit of orthogonality. However, whenever it's necessary, we can

always go back to (4.1) and implement a stochastic search.

As we have discussed in previous section, FB model averaging cannot benefit from the orthogonality, either. To compute (4.5), we must implement a MCMC searching algorithm to sample a sequence of $\gamma$ that converges to the posterior distribution, $\pi(\gamma \mid Y)$.

The potential difficulty with a MCMC procedure here is that it's hard to sample enough to assess the performance when $p$ is large. Even when $p$ is small, it will take a long time to generate a large set of replications to evaluate the performance. Both the sample size of $\gamma$s and the number of replications are important to evaluate the performance of FB model averaging estimator, since the estimator obtained through MCMC is an asymptotic approximation to (4.5). In addition, the multimodality in posterior of $\gamma$ can be another challenge for the MCMC searching algorithm, since the searching can be easily trapped in a local maximum. Such problem may also occur with $C_{CML}$ when (4.1) is attempted for $C_{CML}$.

## 4.4 Simulations

We applied the same data generation procedure used in section 2.4 and section 3.4. The parameters were prespecified as follow: $c = 5$, $\alpha = 1$, $b$=1 and $wa = wb = 1$.

The simulation evaluation consists of four parts: the first part is a comparison of EB model averaging with FB model averaging for $n = p =$

5, the second part is to exam the posteriors and the FB model averaging to gain the insight of the FB procedure, the third part is to compare EB model averaging with FB model averaging for $n = p = 1000$, and the fourth part is to compare the sizes of the models picked by $C_{MML}$, $C_{CML}$ and FB.

Simulation 1: Comparison of EB model averaging with FB model averaging for $p = 5$.

Because the FB model averaging estimator does not have a closed form, we must implement a MCMC algorithm. When $p = 1000$, the MCMC algorithm can only, practically, visit a small portion of the model space and therefore it's hard to evaluate its performance and to compare it with other criteria. For this reason, we choose $p = 5$ so that we will be able to assess the performance through the simulation.

$m = 100$ replications were generated for each set of $q = 0$, 1, 2, 3, 4 and 5. For each set of data generated, $C_{MML}$ model averaging estimators were computed, and a Metropolis-Hastings algorithm was applied to generate a Markov Chain of $\gamma$ for computing the FB model averaging estimator. In the MCMC algorithm, the number of iterations is 5000 and the burn-in period is from 1 to 500 (first 500 were thrown away), i.e., the sample size is 4500. Each time, the MCMC started with a randomly picked single-variable model and then randomly added or removed one variable at a time. The way the MCMC searched in the model space is similar to a random walk. For each model generated, the corresponding conditional posterior mean, $E(\beta_\gamma \mid \hat{\beta}_\gamma, \gamma)$, was computed. The sample mean of 4500 $E(\beta_\gamma \mid \hat{\beta}_\gamma, \gamma)$s was computed to obtain

| q | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| MMLLS | 1.07 | 3.35 | 4.21 | 5.60 | 5.29 | 5.40 |
| CMLLS | 2.69 | 4.16 | 4.53 | 5.37 | 4.98 | 5.18 |
| FBLS | 0.73 | 3.44 | 4.32 | 5.56 | 5.49 | 5.42 |
| MMLPM | 0.29 | 2.30 | 3.38 | 4.62 | 4.74 | 4.84 |
| CMLPM | 0.30 | 2.31 | 3.35 | 4.49 | 4.60 | 4.67 |
| FBPM | 0.15 | 2.31 | 3.30 | 4.55 | 4.99 | 4.96 |
| MMLMA | 0.24 | 2.28 | 3.21 | 4.40 | 4.56 | 4.70 |
| FBMA | 0.03 | 2.22 | 3.23 | 4.98 | 6.21 | 6.60 |

Table 4.1: EB vs FB: Average losses for BLS, BPM and Bayes Model Averaging (BMA) procedures with $p$=5 and $c$=5. m=100, n=5, $\alpha = 1$, $b = 0$ and $wa = wb = 1$.

the FB model averaging estimator.

The average predictive losses are reported in Table (4.1) and Figure (4.1). The model averaging has improved $C_{MML}$ uniformly over its corresponding least-squares and conditional posterior mean estimators. When $q =$ 2, 3 and 4, MMLMA is the best. For other $q$s, although MMLMA does not dominate, it's very close to the best. The FB model averaging achieved excellent performance for q=0, 1 and 2, but performed very poorly for saturated models. This can be very likely caused by the bimodality in the posterior probabilities of $\gamma$ (actually, the posterior probability distribution of $\gamma$ is multimodal. But, if we consider only the subset of highest probable models of each size, the posterior probability presents bimodality). The MCMC searching can be easily trapped in one of the two dominant modes: one towards the null model and one towards the full model, or other local maximum. If the searching starts with a parsimonious model, the Markov Chain may stay for a long time in a neighborhood of the small model whose posterior probability is relatively high, and can hardly move out, and vice versa.

Simulation 2: A closer look of the FB model averaging

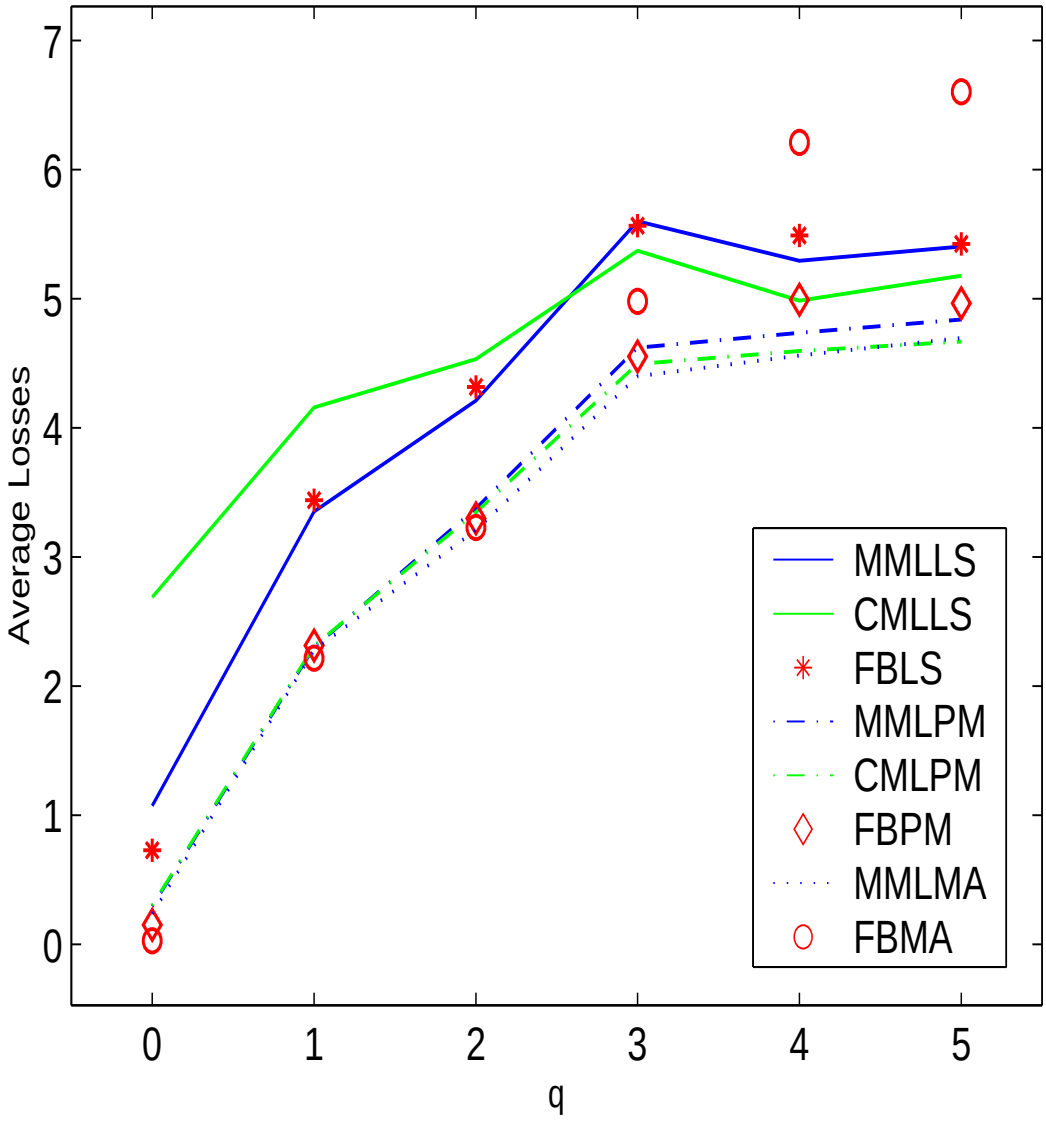


Figure 4.1: EB vs FB: Average losses for BLS, BPM and Bayes Model Averaging (BMA) procedures with $p$=5 and $c$=5. m=100, n=5, $\alpha = 1$, $b = 0$ and $wa = wb = 1$.

The phenomenon discussed above can be well depicted by Figure 4.2 and 4.3. The two plots were obtained from the following simulations: we fixed $n = p = 5$,$c = 5$, $\alpha = 1$, $b = 1$ and $wa = wb = 1$. $m = 5$ replications were generated for $q = 0$, 1, 2, 3, 4 and 5. For each $q$ and each replication, all the posterior probabilities of the 32 models were calculated and plotted in Figure 4.2. Each row contains the five plots corresponding to the five replications for each $q$. The vertical axis is for the posteriors and the horizontal axis is for the index of the models. On the horizontal axis, 1 is the index of the null model; 2 to 6 are the indices of the five one-variable models; 7 to 16, the ten two-variable models, 17 to 26, the ten three-variable models; 27 to 31, the five four-variable models and 32, the full model. The models of the same size are arranged in increasing order of posterior probability from left to right. For each $q$ and each replication, the histogram of the number of variables in the models visited by the MCMC was then plotted in Figure 4.3. Again, the five plots in each row corresponds to the five replications for each $q$.

From Figure 4.2 we can see that, with few exception, the global maximum posterior is usually at either the null model or the full model. However, there are occasions that some other models are almost as equally probable as the null model or the full model, especially when the size of actual model is relatively full. This multimodality together with the overall bimodality in the posterior of $\gamma$ contribute, at least in part, to the poor performance of FB model averaging towards the full model. The reason is that the MCMC can be trapped at the null model or highly probable models near the null model.

For example, when the actual model contains three variables, the posterior probabilities of the "best" 2-variable model, 3-variable model and 4-variable model are all almost the same as that of the full model (see the 4th row of Figure 4.2). But, most of the models visited by the MCMC are 2-variable or smaller models, and only few of them are 3-variable or larger models (see the 4th row of Figure 4.3). Even when the actual model is the full model and the overall trend of the posterior of $\gamma$ is increasing, MCMC has difficulty in reaching the full model or 4-variable model.

On the other hand, FB model averaging has shown its potential in identifying dominant variables. Table 4.2 shows that the actual $\beta$s and FB model averaging estimates (i.e the betas and FBMAs in the table) from the same data used to produce Figure 4.2 and Figure 4.3. In most of the cases, FB model averaging has identified the variables with nonzero coefficients as the most probable variables in terms of their relative magnitudes in each estimate of the actual $\beta$. It is especially significant when $q = 3$. Actually, the potential of FB model averaging can be appreciated when the signal becomes stronger. Simulation with $c = 25$ has shown that FB model averaging delivers better relative performance than it does for the case of $c = 5$ (see Table 4.3 and Figure 4.4). In Table 4.1, FBMA are the worst when $q = 4$ and 5, and its average losses are 1.36 times and 1.40 times as large as that of the best criteria, respectively. But, in Table 4.3, they are only 1.12 and 1.18 times larger. In addition, when $c = 25$, FBMA dominates all the other criteria for $q$=1, 2 and 3 (ties with MMLMA for $q = 3$). However, it's unclear why it becomes worse

for the null model.

If FB is desired, FB model averaging may be potentially better than other FB procedures. There is no doubt that the performance of the FB model averaging procedure can be further improved from the side of MCMC algorithm. However, the tremendous amount of the time taken by the stochastic search and the relatively small portion it can visit when $p$ is large may eventually prevent the FB model averaging procedure from being a practical solution to the variable selection problem. Furthermore, the multimodality (or the overall bimodality) of the posterior may be the fundamental problem of the FB procedure.

Simulation 3: Comparison of EB model averaging with FB model averaging for $p = 1000$.

Although choosing $p$ =5 enables us to assess the performance of FB easier, it may bias the performance of $C_{MML}$ since the sample size of $n$=$p$=5 is too small for a criterion like $C_{MML}$ that heavily relies on the data. In this simulation, we will go back to $p = 1000$, but let $q = 0$, 1, 2, 3, 4, 5, 25, 300, 500, 750 and 1000, so that we are able to see how each criterion dose for both small models and large models.

With the same data generation procedure and all the other parameters being kept the same as before, the average losses are reported in Table 4.4 and are plotted in Figure 4.5. Both the table and the plot show that MMLMA has performed extremely well and significantly dominates all the other criteria

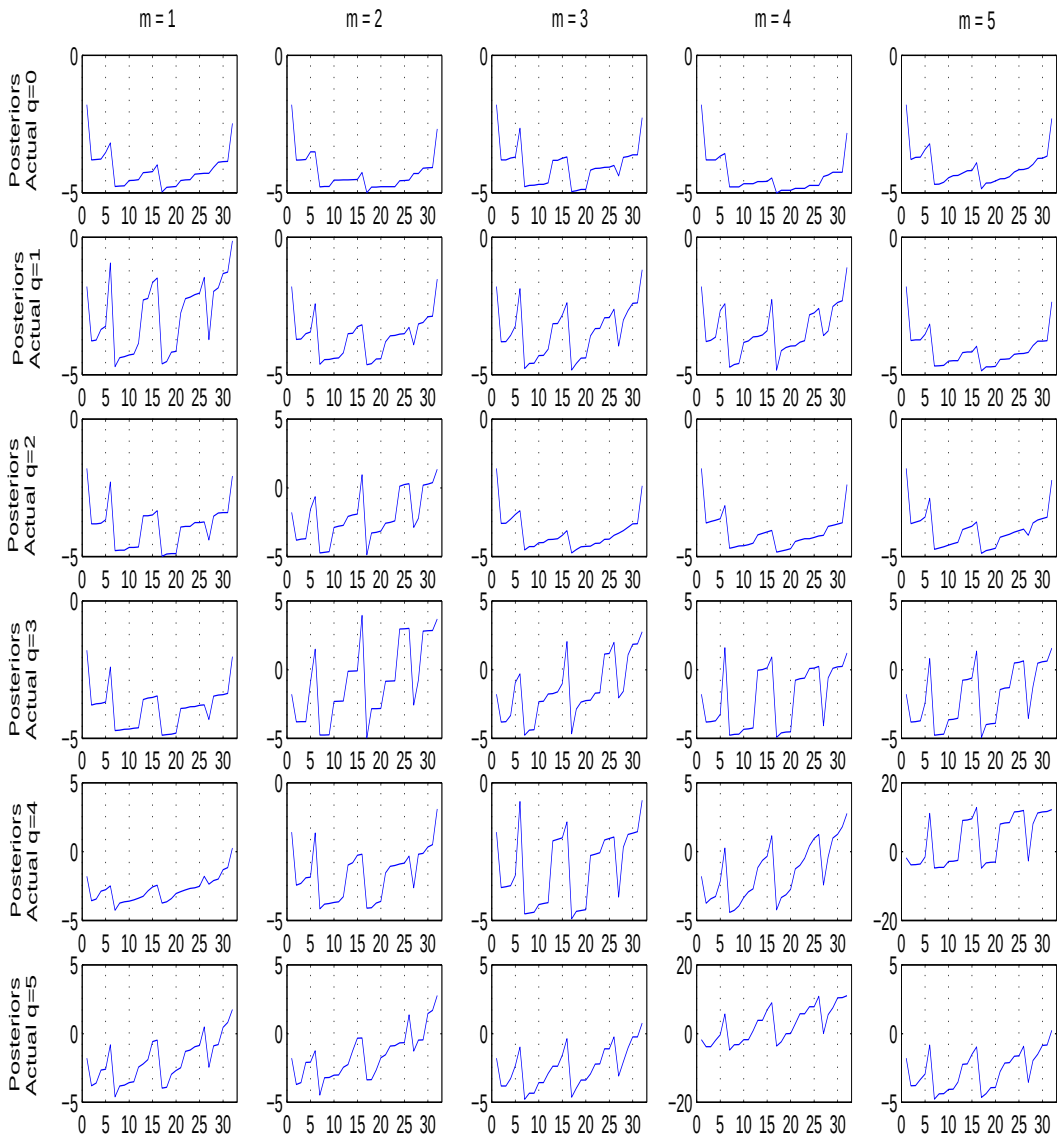


Figure 4.2: Posterior of $\gamma$ (the posterior probabilities are known up to a normalizing constant). m=5, n=p=5, c=5, $\alpha = 1$, $b = 0$ and $wa = wb = 1$. The five plots in each row corresponds to the five replications for each q.

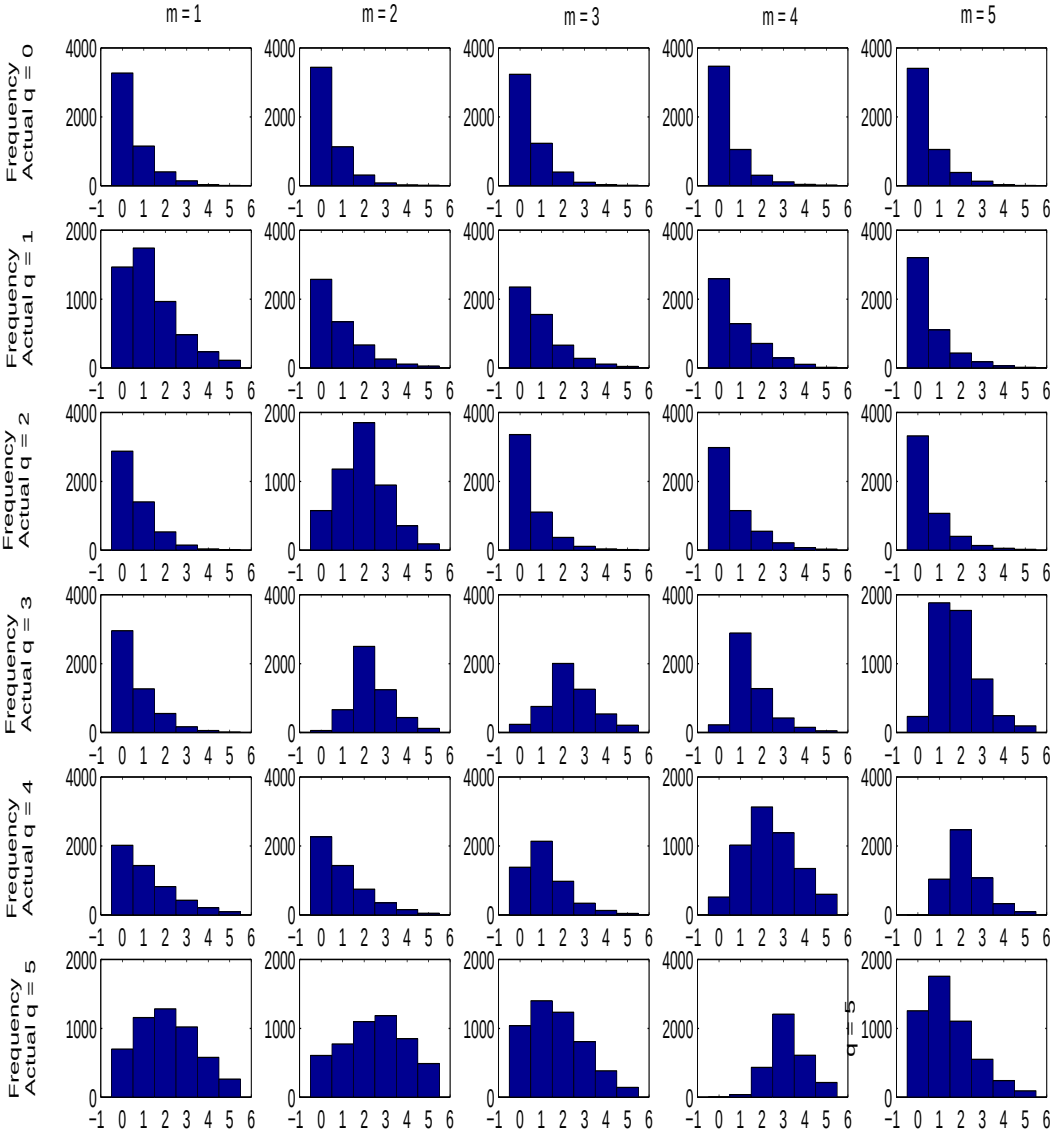


Figure 4.3: Histogram of the size ($q_\gamma$) of the models generated by the MCMC algorithm. m=5, n=p=5, c=5, $\alpha = 1$, $b = 0$ and $wa = wb = 1$. The five plots in each row corresponds to the five replications for each q.

| q=0 | X1 | X2 | X3 | X4 | X5 | q=3 | X1 | X2 | X3 | X4 | X5 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| beta | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | beta | 0.87 | 0.20 | -1.42 | 0.00 | 0.00 |
| FBMA | -0.01 | -0.11 | 0.00 | 0.01 | -0.06 | FBMA | 0.01 | 0.03 | -0.35 | 0.03 | 0.02 |
| beta | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | beta | -1.84 | -0.59 | -2.66 | 0.00 | 0.00 |
| FBMA | 0.05 | 0.06 | 0.00 | 0.01 | 0.00 | FBMA | -3.25 | 0.05 | -2.15 | 0.05 | 0.03 |
| beta | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | beta | 0.05 | -2.24 | -2.12 | 0.00 | 0.00 |
| FBMA | -0.01 | 0.02 | -0.02 | 0.19 | 0.00 | FBMA | -0.05 | -2.32 | -2.00 | 0.33 | 0.01 |
| beta | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | beta | -2.72 | -0.09 | -2.52 | 0.00 | 0.00 |
| FBMA | 0.00 | 0.03 | 0.00 | 0.00 | -0.03 | FBMA | -3.02 | -0.03 | -0.26 | 0.01 | 0.07 |
| beta | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | beta | -2.61 | -1.03 | -0.59 | 0.00 | 0.00 |
| FBMA | 0.01 | -0.07 | 0.02 | 0.09 | -0.02 | FBMA | -2.74 | -0.81 | 0.03 | 0.00 | -0.07 |
| q=1 | | | | | | q=4 | | | | | |
| beta | 1.92 | 0.00 | 0.00 | 0.00 | 0.00 | beta | 1.80 | 0.52 | -2.21 | 3.00 | 0.00 |
| FBMA | 1.24 | -0.26 | -0.17 | 0.05 | -0.03 | FBMA | 0.30 | 0.28 | -0.09 | 0.46 | -0.11 |
| beta | 1.54 | 0.00 | 0.00 | 0.00 | 0.00 | beta | -0.16 | -0.74 | -1.89 | 1.11 | 0.00 |
| FBMA | 0.36 | 0.04 | 0.08 | 0.04 | 0.09 | FBMA | 0.12 | -0.10 | -0.64 | 0.05 | 0.04 |
| beta | -2.69 | 0.00 | 0.00 | 0.00 | 0.00 | beta | -1.91 | -2.69 | -0.27 | -0.15 | 0.00 |
| FBMA | -0.59 | -0.01 | -0.15 | 0.01 | -0.08 | FBMA | -0.17 | -1.43 | -0.03 | -0.04 | 0.00 |
| beta | 3.16 | 0.00 | 0.00 | 0.00 | 0.00 | beta | 3.43 | -1.36 | -3.01 | 1.05 | 0.00 |
| FBMA | 0.37 | 0.03 | 0.01 | -0.06 | -0.29 | FBMA | 1.19 | -0.28 | -2.51 | 0.41 | 0.08 |
| beta | -0.13 | 0.00 | 0.00 | 0.00 | 0.00 | beta | -0.46 | -4.59 | 0.30 | 3.56 | 0.00 |
| FBMA | -0.05 | 0.02 | 0.02 | 0.11 | 0.03 | FBMA | 0.07 | -5.60 | 0.03 | 1.94 | -0.16 |
| q=2 | | | | | | q=5 | | | | | |
| beta | -1.44 | 0.85 | 0.00 | 0.00 | 0.00 | beta | -2.76 | 0.65 | -0.96 | 0.12 | -0.82 |
| FBMA | -0.37 | 0.03 | 0.00 | 0.00 | -0.01 | FBMA | -1.75 | 0.16 | -0.03 | 0.72 | -0.69 |
| beta | 2.45 | -4.19 | 0.00 | 0.00 | 0.00 | beta | -2.29 | 2.32 | -0.87 | -3.09 | 0.71 |
| FBMA | 1.39 | -1.94 | 0.09 | 0.07 | 0.00 | FBMA | -0.13 | 1.64 | 0.21 | -1.13 | 1.12 |
| beta | 1.51 | 1.27 | 0.00 | 0.00 | 0.00 | beta | -0.76 | -2.55 | -0.47 | 2.66 | -2.50 |
| FBMA | 0.06 | 0.03 | -0.01 | -0.08 | -0.01 | FBMA | -0.01 | -1.51 | 0.01 | 0.28 | -0.73 |
| beta | 0.26 | 0.70 | 0.00 | 0.00 | 0.00 | beta | -4.48 | -1.10 | 1.03 | -0.72 | 2.77 |
| FBMA | 0.14 | 0.02 | 0.03 | 0.04 | 0.04 | FBMA | -4.49 | -2.86 | -0.04 | 0.07 | 1.82 |
| beta | -2.22 | 0.47 | 0.00 | 0.00 | 0.00 | beta | 0.85 | 2.11 | -4.74 | -1.44 | -1.57 |
| FBMA | -0.14 | -0.02 | -0.03 | 0.05 | 0.00 | FBMA | -0.01 | 0.34 | -1.43 | -0.18 | -0.03 |

Table 4.2: Comparisons of actual $\beta$(beta) and FB model averaging estimator of $\beta$(FBMA): n=p=5, c=5, $\alpha = 1$, $b = 0$ and $wa = wb = 1$. For each $q$, m=5 replications were generated. The five pair rows of beta and FBMA under the $X$s correspond to the five replications.

| q | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| MMLLS | 0.10 | 2.88 | 4.29 | 5.90 | 5.31 | 5.26 |
| CMLLS | 2.69 | 3.41 | 4.44 | 5.87 | 5.02 | 5.12 |
| FBLS | 0.73 | 3.29 | 4.17 | 5.63 | 5.29 | 5.30 |
| MMLPM | 0.03 | 2.54 | 4.14 | 5.71 | 5.23 | 5.13 |
| CMLPM | 0.30 | 2.54 | 4.15 | 5.70 | 4.98 | 4.99 |
| FBPM | 0.15 | 2.73 | 4.11 | 5.60 | 5.38 | 5.27 |
| MMLMA | 0.03 | 2.40 | 3.81 | 5.26 | 5.05 | 5.06 |
| FBMA | 0.11 | 1.95 | 3.45 | 5.26 | 5.62 | 5.89 |

Table 4.3: EB vs FB: Average losses for BLS, BPM and Bayes Model Averaging (BMA) procedures with $p = 5$ and $c = 25$. m=100, n=5, $\alpha = 1$, $b = 0$ and $wa = wb = 1$.

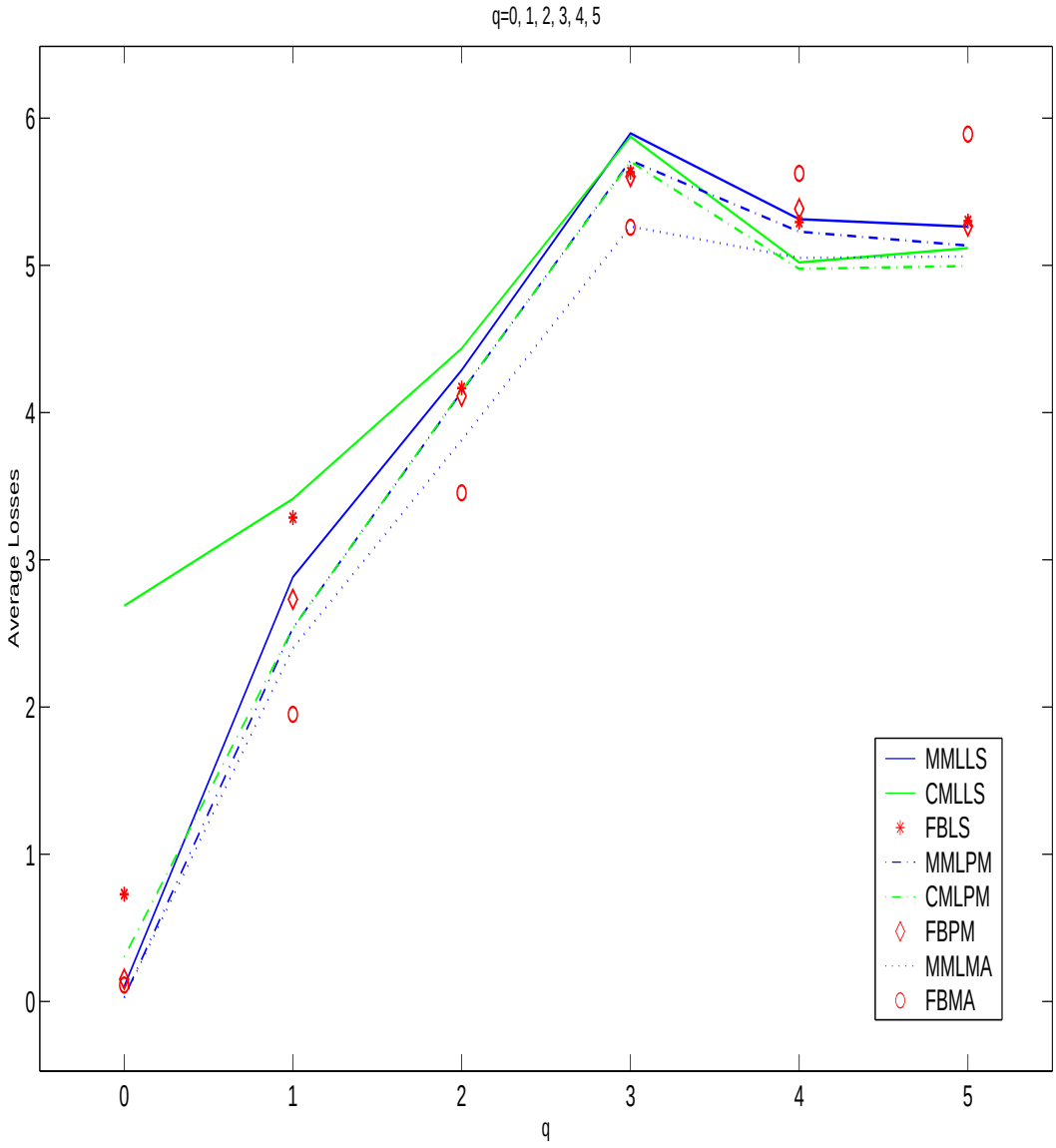


Figure 4.4: EB vs FB: Average losses for BLS, BPM and Bayes Model Averaging (BMA) procedures with $p$=5 and $c$=25. m=100, n=5, $\alpha = 1$, $b = 0$ and $wa = wb = 1$.

| q | 0 | 1 | 2 | 3 | 4 | 5 | 25 | 300 | 500 | 750 | 1000 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| MMLLS | 5.83 | 9.32 | 10.09 | 14.87 | 18.09 | 21.10 | 74.16 | 621.88 | 892.22 | 992.35 | 991.26 |
| CMLLS | 0.00 | 4.12 | 7.29 | 12.48 | 14.47 | 19.34 | 77.72 | 683.27 | 986.92 | 1240.32 | 1178.71 |
| FBLS | 0.00 | 4.12 | 7.45 | 12.68 | 14.87 | 19.48 | 77.34 | 682.78 | 986.41 | 1239.88 | 1178.33 |
| MMLPM | 1.87 | 5.12 | 7.70 | 11.47 | 14.33 | 17.48 | 66.66 | 522.59 | 723.08 | 792.44 | 826.84 |
| CMLPM | 0.00 | 4.02 | 7.22 | 12.41 | 14.31 | 19.24 | 76.73 | 660.09 | 940.13 | 1160.09 | 1044.12 |
| FBPM | 0.00 | 3.87 | 7.30 | 12.52 | 14.54 | 19.28 | 75.86 | 659.04 | 939.05 | 1159.17 | 1043.51 |
| MMLMA | 1.95 | 5.24 | 7.36 | 11.76 | 13.79 | 17.49 | 59.66 | 438.71 | 617.34 | 754.01 | 827.36 |
| FBMA | 0.09 | 3.91 | 7.30 | 13.05 | 14.82 | 19.05 | 80.16 | 714.44 | 1065.09 | 1446.38 | 1767.44 |

Table 4.4: EB vs FB: Average losses for BLS, BPM and Bayes Model Averaging (BMA) procedures with $p = 1000$ and $c = 5$. m=100, n=1000, $\alpha = 1$, $b = 0$ and $wa = wb = 1$.

for $q = 25$ and larger. When actual model is small ($q = 0$, 1, 2, 3, 4 and 5), MMLMA does not perform as well as it does for large models, and MMLLS is off even more. We will discuss such phenomenon further in the next simulation.

When $q$ is small ($q < 25$), FBMA delivers fine performances, although it's not improved as much as it can be. However, as $q$ increases, FBMA can be very wild. For example, its average losses for $q$= 750 and 1000 have been off way too much. As we expected, FBMA is greatly disadvantaged by the huge model space – it's practically difficult to run the MCMC long enough to search the entire space. Besides, the bimodality can be fatal to FBMA in this case.

Simulation 4: Comparison of the sizes of models picked by $C_{MML}$, $C_{CML}$ and $C_{FB}$.

Due to the assumption that both the $\beta$ and the error come from normal distributions with mean zero but different covariance matrices, all the criteria eventually choose a model that is smaller than the actual model. However, among the three criteria, $C_{MML}$ tends to pick larger models than $C_{CML}$ or $C_{FB}$ does. Such feature enables $C_{MML}$ to capture more variables that are in the

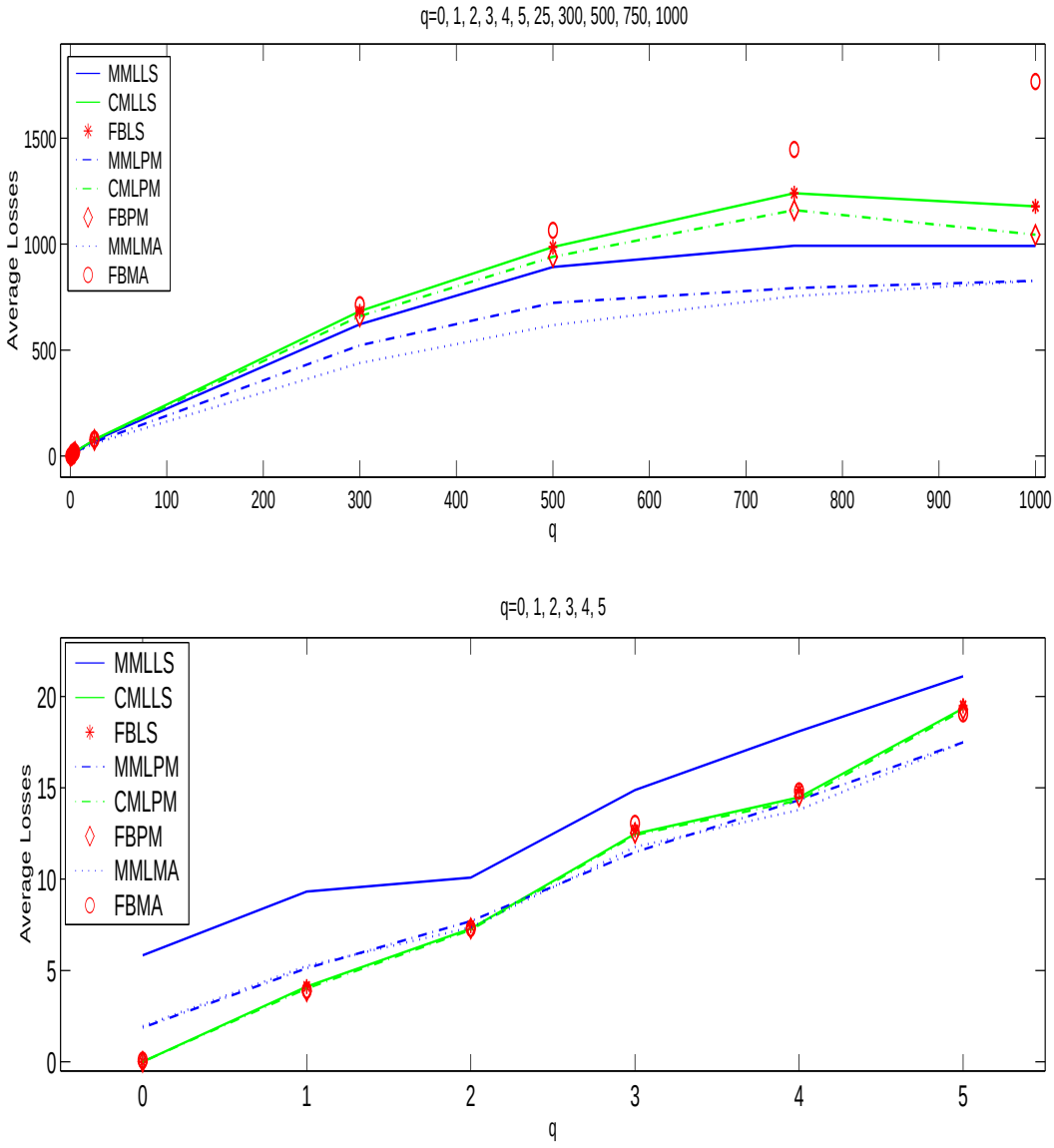


Figure 4.5: EB vs FB: Average losses for BLS, BPM and Bayes Model Averaging (BMA) procedures with $p = 1000$ and $c = 5$. m=100, n=1000, $\alpha = 1$, $b = 0$ and $wa = wb = 1$.

actual model and to gain more predictive power. Table 4.5 shows that while $C_{MML}$, $C_{CML}$ and $C_{FB}$ all pick smaller models than the actual model, $C_{MML}$ tends to pick larger models and, hence, a model closer to the actual model. For example, for the first replication (m = 1), when the size of actual model is $q = 25$, $C_{MML}$ picks model of size 7, both $C_{CML}$ and $C_{FB}$ pick models of size 4; when the actual model size is 300, both $C_{CML}$ and $C_{FB}$ pick models of size 82. Such pattern is seen in all the 12 replications. The draw back of this nice feature is that when the actual model is small, $C_{MML}$ may overestimate the model, i.e., choose a model that is bigger than the actual model. For example, when $m = 2$ and $q = 4$, the model picked by $C_{MML}$ is of size 22; $m = 7$, $q = 0$, the model is of size 6. This may be the reason why $C_{CML}$ and $C_{FB}$ presents better performance than $C_{MML}$ does for small $q$s as we can see from Table 3.1 or 4.4. However, the reversed performance for $C_{MML}$ presented in Table 4.1 or 4.3 may be due to the lack of enough sample – since we choose $n = p = 5$.

| q | 0 | 1 | 2 | 3 | 4 | 5 | 25 | 300 | 500 | 750 | 1000 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| data 1 | | | | | | | | | | | |
| $C_{MML}$ | 0 | 0 | 0 | 0 | 1 | 0 | 7 | 198 | 336 | 1000 | 1000 |
| $C_{CML}$ | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 82 | 155 | 271 | 1000 |
| $C_{FB}$ | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 82 | 156 | 271 | 1000 |
| data 2 | | | | | | | | | | | |
| $C_{MML}$ | 0 | 0 | 0 | 0 | 22 | 2 | 10 | 142 | 342 | 1000 | 1000 |
| $C_{CML}$ | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 68 | 143 | 275 | 1000 |
| $C_{FB}$ | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 69 | 143 | 275 | 1000 |
| data 3 | | | | | | | | | | | |
| $C_{MML}$ | 0 | 1 | 0 | 1 | 1 | 1 | 7 | 127 | 362 | 1000 | 1000 |
| $C_{CML}$ | 0 | 0 | 0 | 0 | 1 | 0 | 2 | 66 | 169 | 315 | 479 |
| $C_{FB}$ | 0 | 0 | 0 | 0 | 1 | 0 | 2 | 66 | 169 | 315 | 479 |
| data 4 | | | | | | | | | | | |
| $C_{MML}$ | 1 | 0 | 1 | 2 | 0 | 2 | 8 | 151 | 342 | 1000 | 1000 |
| $C_{CML}$ | 0 | 0 | 1 | 0 | 0 | 0 | 6 | 58 | 147 | 310 | 531 |
| $C_{FB}$ | 0 | 0 | 1 | 0 | 0 | 0 | 6 | 58 | 148 | 310 | 531 |
| data 5 | | | | | | | | | | | |
| $C_{MML}$ | 0 | 1 | 0 | 2 | 2 | 1 | 5 | 185 | 303 | 1000 | 1000 |
| $C_{CML}$ | 0 | 0 | 0 | 0 | 2 | 0 | 5 | 85 | 127 | 279 | 454 |
| $C_{FB}$ | 0 | 0 | 0 | 0 | 2 | 0 | 5 | 85 | 127 | 279 | 454 |
| data 6 | | | | | | | | | | | |
| $C_{MML}$ | 0 | 0 | 0 | 0 | 1 | 0 | 12 | 131 | 304 | 1000 | 1000 |
| $C_{CML}$ | 0 | 0 | 0 | 0 | 1 | 0 | 7 | 71 | 176 | 314 | 1000 |
| $C_{FB}$ | 0 | 0 | 0 | 0 | 1 | 0 | 7 | 72 | 176 | 314 | 1000 |
| data 7 | | | | | | | | | | | |
| $C_{MML}$ | 6 | 1 | 0 | 0 | 0 | 1 | 2 | 139 | 356 | 701 | 1000 |
| $C_{CML}$ | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 70 | 157 | 261 | 555 |
| $C_{FB}$ | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 70 | 157 | 261 | 555 |
| data 8 | | | | | | | | | | | |
| $C_{MML}$ | 0 | 0 | 1 | 0 | 0 | 0 | 10 | 163 | 283 | 1000 | 1000 |
| $C_{CML}$ | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 62 | 138 | 264 | 482 |
| $C_{FB}$ | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 62 | 138 | 264 | 482 |
| data 9 | | | | | | | | | | | |
| $C_{MML}$ | 0 | 0 | 0 | 3 | 0 | 0 | 2 | 157 | 333 | 1000 | 1000 |
| $C_{CML}$ | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 58 | 131 | 263 | 541 |
| $C_{FB}$ | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 58 | 131 | 263 | 541 |
| data 10 | | | | | | | | | | | |
| $C_{MML}$ | 0 | 1 | 0 | 1 | 1 | 4 | 5 | 160 | 317 | 882 | 1000 |
| $C_{CML}$ | 0 | 0 | 0 | 0 | 1 | 0 | 3 | 67 | 148 | 287 | 1000 |
| $C_{FB}$ | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 67 | 148 | 288 | 1000 |
| data 11 | | | | | | | | | | | |
| $C_{MML}$ | 0 | 1 | 1 | 0 | 1 | 1 | 8 | 135 | 371 | 1000 | 1000 |
| $C_{CML}$ | 0 | 0 | 1 | 0 | 1 | 1 | 3 | 77 | 148 | 381 | 478 |
| $C_{FB}$ | 0 | 0 | 1 | 0 | 1 | 1 | 3 | 77 | 148 | 381 | 478 |
| data 12 | | | | | | | | | | | |
| $C_{MML}$ | 0 | 5 | 0 | 0 | 3 | 0 | 6 | 175 | 354 | 594 | 1000 |
| $C_{CML}$ | 0 | 0 | 0 | 0 | 2 | 0 | 3 | 81 | 138 | 236 | 551 |
| $C_{FB}$ | 0 | 0 | 0 | 0 | 2 | 0 | 3 | 81 | 138 | 239 | 551 |

Table 4.5: Comparison of the sizes of the models selected by $C_{MML}$, $C_{CML}$ and $C_{FB}$: m=12, n=p=1000, c=5, $\alpha = 1$, $b = 0$ and $wa = wb = 1$.

# Chapter 5

# Discussion

In this research, we have investigated the potential of FB methods for the variable selection problem. Various priors have been considered and discussed. Both EB and FB procedures followed by least-squares and posterior-mean estimators of the coefficients were intensively investigated. Model averaging as an alternative to selecting one single model has also been explored. In all cases, $C_{MML}$ has outperformed FB surprisingly well in the spite of its known inadmissibility. By estimating $c$ and $\omega$ from the marginal likelihood, $C_{MML}$ successfully incorporates all the information about the model contained in the data into the selection criterion. The interesting observation that $C_{MML}$ tends to pick larger models also offers an intuitive explanation of its excellent performance over $C_{CML}$ and $C_{FB}$, and may be worth further exploration.

The FB procedures, which are promising theoretically, did not deliver better performance than the EB procedures. The bimodality has been shown to be a central problem. Although FB model averaging has shown its potential for improvement, it still suffers from the bimodality in the posterior probabilities of the models, and, in addition, the difficulty of MCMC to search in a huge model space.

In addition, as we have discussed before, the specification of the prior and the choice of selection rule based on the posterior are critical. As we have seen, a noninformative prior does not necessarily lead to a desirable FB procedure; although the posterior mode is admissible under a 0-1 loss, it may be poor under squared-error loss. While choosing an appropriate prior is a necessity for the FB approach, it is completely avoided by the EB approach. Other procedures that avoid prior choice include the Fractional Bayes Factor of O'Hagan (1995), the intrinsic Bayes factor of Berger and Pericchi (1996), and the predictive criteria of Laud and Ibrahim (1995). The problem of finding appropriate priors for FB variable selection warrants further exploration. As for the selection rule, the posterior mode may not be appropriate under the squared-error loss. Recent research by Barbieri and Berger (2002) found that, in the context of normal linear model selection and under certain conditions, the optimal model was not the posterior mode, but the median probability model which they define as "the model consisting of those variables which have overall posterior probability greater than or equal to 1/2 of being in a model."

Another direction for future exploration is to extend this research to the nonorthogonal case. In this research, we have mainly focused on the orthogonal case (i.e., orthogonal $X$). The orthogonality has simplified the computation of both the EB posterior and the FB posterior. It also enables us to compute the $C_{MML}$ estimator of $c$ and $\omega$ and $C_{MML}$ model averaging estimator of $\beta$ straightforwardly. If $X$ is nonorthogonal, $C_{MML}$ will be ruled out. $C_{CML}$

and $C_{FB}$ can still be applied in all the three procedures. But, finding the posterior mode won't be straightforward since it's impractical to evaluate all the $2^p$ models. A stepwise method can be applied to reduce the sample space to a promising subset with respect some criterion. George and Foster (2000) studied the nonorthogonal case for $C_{CML}$ and other fixed penalty criteria.

Throughout this dissertation, we have assumed $\sigma$ to be known and have set it to 1 in our simulations. By doing so, we can investigate the influence of the two approaches to hyperparameter uncertainty (estimating them from the data and putting priors on them) more easily without worrying about the impact from estimating or integrating $\sigma$. However, $\sigma$ can be incorporated into the procedures. One straightforward way is to replace it with an estimate. For example, one can use the mean residual sum of squares of the full model $(Y'Y - SS_p)/(n-p)$, which is an unbiased estimator of $\sigma$. When $n = p$, as in the wavelet context, one can use $\hat{\sigma} = \text{median}(|\hat{\beta}_i|)/0.6745$ proposed by Donoho et al. (1995). Another common approach is to introduce a prior on $\sigma$ or $\sigma^2$. An Inverse Gamma, $IG(\nu/2, \nu\lambda/2)$ on $\sigma^2$ has been considered by Clyde, Desimone and Parmigiani (1996), Garthwaite and Dickey (1992), George and McCulloch (1997), Raftery, Madigan and Hoeting (1997) and Chipman, George and McCulloch (2001) and many others. Such a prior can be easily incorporated into the EB formulation and $\sigma^2$ can be easily integrated out given $c$ and $\omega$. In the FB setup, the posterior of $\gamma$ will not have a closed form with this prior on $\sigma^2$ together with the priors we chose for $c$ and $\omega$. However, one can apply a MCMC algorithm to simulate $\gamma$ and $\sigma$ from $p(\gamma|Y,\sigma)$ and $p(\sigma|Y,\gamma)$ suc-

cessively, while keeping in mind that the stochastic search may only visit a very small portion of the model space and may be disturbed in large by the bimodality.

This research, rather than being conclusive in its comparison of EB with FB approaches to variable selection, has served to open up many more questions for future research.

# Appendix

Theorem 2.2.1: Consider the variable selection problem for the linear model (1.1). Suppose the priors of $\beta_\gamma$ and $\gamma$ are (2.3) and (2.4), respectively. Then the conditional Jeffreys prior on $c$ given $\gamma$ is

$$\pi(c \mid \gamma) = \frac{\sqrt{\frac{q_\gamma}{2}}}{1+c}.$$

Proof: Let $\hat{\beta}_\gamma$ be the least-square estimate of $\beta_\gamma$. Since it's a sufficient statistic, the likelihood function $f(y \mid \beta_\gamma, \gamma)$ can be decomposed into a product as $f(y \mid \beta_\gamma, \gamma) = g(y \mid \hat{\beta}_\gamma) \cdot p_{\hat{\beta}_\gamma}(\hat{\beta}_\gamma \mid \beta_\gamma)$. From (1.1), we know that $Y \mid \beta_\gamma, \gamma \sim N(X\beta, \sigma^2 I)$, i.e,

$$f(y \mid \beta_\gamma, \gamma) \;=\; (2\pi)^{-\frac{n}{2}} |I_n \sigma^2|^{-\frac{1}{2}} \exp\left\{ -\frac{(Y - X_\gamma\beta_\gamma)'(Y - X_\gamma\beta_\gamma)}{2\sigma^2} \right\}.$$

Let

$$Y - X_\gamma\beta_\gamma = Y - X_\gamma\hat{\beta}_\gamma + X_\gamma\hat{\beta}_\gamma - X_\gamma\beta_\gamma.$$

Then $f(y \mid \beta_\gamma, \gamma) = g(y \mid \hat{\beta}_\gamma) \cdot p_{\hat{\beta}_\gamma}(\hat{\beta}_\gamma \mid \beta_\gamma)$, where

$$\begin{aligned} g(y \mid \hat{\beta}_\gamma, \gamma) \;&=\; |(X_\gamma' X_\gamma)^{-1}|^{\frac{1}{2}} (2\pi)^{-\frac{n-q_\gamma}{2}} (\sigma^2)^{-\frac{n-q_\gamma}{2}} \\ &\quad \cdot \exp\{ -\frac{(Y - X_\gamma\hat{\beta}_\gamma)'(Y - X_\gamma\hat{\beta}_\gamma)}{2\sigma^2} \} \end{aligned}$$

and $p(\hat{\beta}_\gamma \mid \beta_\gamma, \gamma)$ is the density function of $N_{q_\gamma}(\beta_\gamma, (X_\gamma' X_\gamma)^{-1}\, \sigma^2)$.

$$p(\hat{\beta}_\gamma \mid \beta_\gamma, \gamma) = (2\pi)^{-\frac{q_\gamma}{2}} |(X_\gamma' X_\gamma)^{-1}\sigma^2|^{-\frac{1}{2}} \exp\{ -\frac{(\hat{\beta}_\gamma - \beta_\gamma)'({X_\gamma}' X_\gamma)(\hat{\beta}_\gamma - \beta_\gamma)}{2\sigma^2} \}$$

is the density function of $N_{q_\gamma}(\beta_\gamma, (X'_\gamma X_\gamma)^{-1}\,\sigma^2)$. Now, we have

$$\begin{aligned}\hat{\beta}_\gamma \mid \beta_\gamma, \gamma &\sim N_{q_\gamma}(\beta_\gamma, (X'_\gamma X_\gamma)^{-1}\sigma^2),\\ \beta_\gamma \mid \gamma, c &\sim N_{q_\gamma}(0, c\sigma^2\,(X'_\gamma X_\gamma)^{-1}).\end{aligned}$$

It can be easily shown that

$$\hat{\beta}_\gamma \mid c, \gamma \sim N_{q_\gamma}(0, (1+c)(X'_\gamma X_\gamma)^{-1}\sigma^2).$$

The log likelihood function, $L(c \mid \gamma)$ is then

$$\begin{aligned}L(c \mid \gamma) &= Log(p(\hat{\beta}_\gamma \mid c, \gamma)) \propto -\frac{q_\gamma}{2}\log(1+c) - \frac{\hat{\beta}'_\gamma (X'X)\hat{\beta}_\gamma}{(1+c)\sigma^2}\\ \frac{\partial^2 L}{\partial c^2} &\propto \frac{q_\gamma/2}{(1+c)^2} - \frac{\hat{\beta}'_\gamma (X'X)\hat{\beta}_\gamma}{(1+c)^3\sigma^2}\\ E\left[\frac{\partial^2 L}{\partial c^2}\right] &= \frac{E\left(\hat{\beta}'_\gamma (X'X)\hat{\beta}_\gamma\right)}{(1+c)^3\sigma^2}.\end{aligned}$$

Since

$$\begin{aligned}E\left(\hat{\beta}_\gamma (X'X)^{-1}\hat{\beta}_\gamma\right) &= trace\left((X'X)(1+c)(X'X)^{-1}\sigma^2\right) + E\left(\hat{\beta}_\gamma\right)(X'X)E\left(\hat{\beta}_\gamma\right)\\ &= (1+c)q_\gamma\sigma^2\end{aligned}$$

Therefore,

$$\pi(c \mid \gamma) = \left(-E\left[\frac{\partial^2 L}{\partial c^2}\right]\right)^{\frac{1}{2}} = \left(\frac{q_\gamma/2}{(1+c)^2}\right)^{\frac{1}{2}} = \frac{\sqrt{q_\gamma/2}}{1+c}.$$

# Bibliography

[1] H. Akaike (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In B.N. Petrov and F. Csaki, editors, *2nd International Symposium on Information Theory*, pages 267–81. 1973.

[2] H. Akaike (1978). A New Look at the Bayes Procedure. *Biometrika*, 65(1):53–59, 1978.

[3] M.M. Barbieri and J.O. Berger (2002). Optimal Predictive Model Selection. ISDS discussion paper.). Technical Report 02-02, Institute of Statistics & Decision Sciences, Duke University, Durham, NC 27708-0251,USA, 2002.

[4] J.O. Berger (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, 1985.

[5] J.O. Berger and L.R. Pericchi (1996). The Intrinsic Bayes Factor for Model Selection and Prediction. *Journal of the American Statistical Association*, 91(433):109–122, 1996.

[6] K.N. Berk (1978). Comparing Subset Regression Procedures. *Technometrics*, 20:1–6, 1978.

[7] J. M. Bernardo (1979). Reference Posterior Distributions for Bayesian Inference. *Journal of the Royal Statistical Society, series B*, 41(2):113–147, 1979.

[8] G.E.P. Box and G.C. Tiao (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley, 1973.

[9] H. Chipman, E.I. George, and R. McCulloch (2001). The Practical Implementation of Bayesian Model Selection (with discussion). In P. Lahiri, editor, *In Model Selection, IMS Lecture Notes – Monograph Series*, Volume 38, pages 65–134. 2001.

[10] M. Clyde, H. Desimone, and G. Parmigian (1996). Prediction Via Orthogonalized Model Mixing. *Journal of the American Statistical Association*, 91(435):1197–1208, 1996.

[11] M. Clyde and E.I. George (1999). Empirical Bayes Estimation in Wavelet Nonparametric Regression. In P. Muller and B. Vidakovic, editors, *Bayesian Inference in Wavelet-Based Models*, pages 309–322. Springer-Verlag, 1999.

[12] M. Clyde and E.I. George (2000). Flexible Empirical Bayes Estimation for Wavelets. *Journal of the Royal Statistical Society, series B*, 62(4):681–698, 2000.

[13] M. Clyde, G. Parmigian, and B. Vidakovic (1998). Multiple Shrinkage and Subset Selection in Wavelets. *Biometrika*, 85(2):391–401, 1998.

[14] J.J. Deely and D.V. Lindley (1981). Bayes Empirical Bayes. *Journal of the American Statistical Association*, 76(376):833–841, 1981.

[15] D. Donoho, I. Johnstone, G. Kerkyacharian, and D. Picard (1995). Wavelet Shrinkage: Asymptopia? *Journal of the Royal Statistical Society, series B*, 57(2):301–369, 1995.

[16] D. Edwards and T. Havranek (1987). A Fast Model Selection Procedure for Large Families of Models. *Journal of the American Statistical Association*, 82(397):205–213, 1987.

[17] M. A. Efroymson (1960). Multiple regression analysis. In A. Ralston and H. S. Wilf, editors, *Mathematical methods for digital computers*, pages 191–203. New York: Wiley, 1960.

[18] D.P. Foster and E.I. George (1994). The Risk Inflation Criterion for Multiple Regression. *Annals of Statistics*, 22(4):1947–1975, 1994.

[19] G.M. Furnival and R.W.J. Wilson (1974). Regression By Leaps and Bounds. *Technometrics*, 16:499–511, 1974.

[20] P.H. Garthwaite and J.M. Dickey (1992). Elicitation of Prior Distributions for Variable-Selection Problems in Regression *Annals of Statistics*, 20(4):1697–1719, 1992.

[21] S. Geisser (1979). "Discussion" of Bernardo's "Reference Posterior Distributions for Bayesian Inference". *Journal of the Royal Statistical Society, series B*, 41(2):136–137, 1979.

[22] S. Geisser (1984). On Prior Distribution for Binary Trials (with discussions). *The American Statistician*, 38(4):244–251, 1984.

[23] E.I. George (2000). The Variable Selection Problem. *Journal of the American Statistical Association*, 95(452):1304–1308, 2000.

[24] E.I. George and D.P. Foster (2000). Calibration and Empirical Bayes Variable Selection. *Biometrika*, 87(4):731–747, 2000.

[25] E.I. George and R. McCulloch (1993). Variable Selection Via Gibbs Sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

[26] E.I. George and R. McCulloch (1997). Approaches for Bayesian Variable Selection. *Statistica Sinica*, 7:339–374, 1997.

[27] R.R. Hocking (1976). The Analysis and Selection of Variables in Linear Regression. *Biometrics*, 32(1):1–49, 1976.

[28] R.R. Hocking and R.N. Leslie (1967). Selection of the Best Subset in Regression Analysis. *Technometrics*, 9:531–540, 1967.

[29] H. Jeffreys (1961). *Theory of Probability (3rd edn.)*. Oxford University Press, 1961.

[30] P.W. Laud and J.G. Ibrahim (1996). Predictive Specification of Prior Model Probabilities in Variable selection. *Biometrika*, 83(2):267–274, 1996.

[31] C.L. Mallows (1973). Some Comments On $C_p$. *Technometrics*, 15:661–671, 1973.

[32] C.L. Mallows (1995). More Comments On $C_p$. *Technometrics*, 37:362–372, 1995.

[33] A. Miller (1984). Selection of Subsets of Regression Variables(with discussion). *Journal of the Royal Statistical Society, series A*, 147(3):389–425, 1984.

[34] A. Miller (1990). *Subset Selection in Regression.* Chapman and Hall, 1990.

[35] P.M. Narendra and K. Fukunaga (1977). A Branch and Bound Algorithm for Feature Subset Selection. *IEEE Transactions on Computers*, 26(9):917–922, 1977.

[36] M.R. Novick and W.J. Hall (1965). A Bayesian Indifference Procedure. *Journal of the American Statistical Association*, 60(312):1104–1117, 1965.

[37] A. OHagan (1995). Fractional Bayes Factors for Model Comparison (with discussion). *Journal of the Royal Statistics Society, Series B*, 57(1):99–138, 1995.

[38] A.E. Raftery, D. Madigan, and J.A. Hoeting (1997). Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association*, 92(437):179191, 1997.

[39] G. Schwarz (1978). Estimating the Dimension of a Model. *Annals of Statistics*, 6(2):461–464, 1978.

# Vita

Wen Cui was born in Guizhou, China on August 15, 1968, the daughter of Caiyun Liu and Jingbo Cui. From September, 1986 to July 1990, she studied Mathematical Statistics in East China Normal University, Shanghai, China. After she graduated from East China Normal University with B.S. in July, 1990, She worked as a statistician in the 11th construction company of China National Nonferrous Metals Industry Corporation, Liuzhou, GuangXi, China. From January 1994 to August 1995, she worked as an account manager in China Resource Machinery (Beijing), the agent of Computer Associate International, Inc., Beijing, China. In Fall, 1995, she entered the graduate program of Department of Mathematics and Statistics at Stephen F. Austin State University, Nacogdoches, Texas, and graduated with M.S. in May, 1997. During the summer of 1997, she worked as an adjunct faculty in Department of Mathematics and Statistics at Stephen F. Austin State University. From August, 1997 to May, 1998, she attended Southern Methodist University, Dallas, Texas. In August, 1998, she entered the Ph.D program of Department of Management Science and Information Systems at University of Texas at Austin. During the years of her graduate studies, she also worked as a Teaching assistant, Research Assistant, Assistant Instructor at University of Texas. In Spring 1999, she worked as an intern modeler at Zilliant Inc., Austin, TX, and in summer 2000, she worked as an intern statistician at Dell, Austin, TX.

Permanent address: 15850 Garrison Circle
Austin, Texas 78717

This dissertation was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.